

Visual Analytics Methodologies on Causality Analysis

by

Hong Wang

A Dissertation Presented in Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy

Approved July 2019 by the  
Graduate Supervisory Committee:

Ross Maciejewski, Chair  
Jingrui He  
Hasan Davulcu  
Cameron Thies

ARIZONA STATE UNIVERSITY

August 2019

## ABSTRACT

Causality analysis is the process of identifying cause-effect relationships among variables. This process is challenging because causal relationships cannot be tested solely based on statistical indicators as additional information is always needed to reduce the ambiguity caused by factors beyond those covered by the statistical test. Traditionally, controlled experiments are carried out to identify causal relationships, but recently there is a growing interest in causality analysis with observational data due to the increasing availability of data and tools. This type of analysis will often involve automatic algorithms that extract causal relations from large amounts of data and rely on expert judgment to scrutinize and verify the relations. Over-reliance on these automatic algorithms is dangerous because models trained on observational data are susceptible to bias that can be difficult to spot even with expert oversight. Visualization has proven to be effective at bridging the gap between human experts and statistical models by enabling an interactive exploration and manipulation of the data and models. This thesis develops a visual analytics framework to support the interaction between human experts and automatic models in causality analysis. Three case studies were conducted to demonstrate the application of the visual analytics framework in which feature engineering, insight generation, correlation analysis, and causality inspections were showcased.

*To my future children*

## ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude towards my PhD advisor, Dr. Ross Maciejewski, for his support and guidance on my research. I would like to thank my wife (and colleague) for her significant contributions to my research and personal well being. I also would like to thank Arizona State University for providing an opportunity for me to continue learning and enhancing myself, and I have no regret for my time spent here.

This work is supported by the U.S. Department of Homeland Security under Award Number, 2017-ST-061-QA0001 and the National Science Foundation, Grant No. 1350573. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

## TABLE OF CONTENTS

	Page
LIST OF TABLES .....	vii
LIST OF FIGURES .....	vii
CHAPTER	
1 INTRODUCTION .....	1
2 BACKGROUND AND RELATED WORK .....	7
2.1 Correlation Visualizations .....	8
2.2 Causality Analysis .....	10
2.2.1 Probabilistic Causation .....	10
2.2.2 Bayesian Networks .....	14
2.2.3 Visual Analytics on Causality Analysis .....	16
3 VISUAL ANALYTICS FRAMEWORK FOR CAUSALITY ANALYSIS .	19
4 IDENTIFYING TOPIC DRIVERS IN MEDIA EVENTS .....	25
4.1 Analytic Tasks .....	28
4.2 Semantic Event Retrieval .....	29
4.2.1 Semantic Lexical Match .....	29
4.2.2 Semantic Similarity Update .....	33
4.2.3 Semantic Interaction for Concept Word Clustering .....	34
4.2.4 Word Weights Update .....	40
4.3 Causality Detection .....	42
4.4 Annotation .....	44
4.4.1 Timeline with Events and Causalities .....	44
4.4.2 Annotating with Text Information .....	46
4.4.3 Bipartite View .....	47
4.4.4 Detail List .....	48

CHAPTER	Page
4.4.5 Entity Annotations .....	49
4.5 Usage Scenario .....	50
4.5.1 Datasets .....	50
4.5.2 Climate-induced Unrest During Drought .....	51
4.5.3 Food Insecurity and Climate Change Media .....	52
4.5.4 User Feedback .....	56
4.6 Discussion .....	58
5 SPATIOTEMPORAL TRADE NETWORK ANALYSIS .....	60
5.1 Global Trade Analytics .....	64
5.1.1 Data Description .....	64
5.1.2 Design Requirements .....	65
5.1.3 Triadic Analysis .....	68
5.1.4 Dependency and Leverage .....	70
5.1.5 Clustering Coefficient .....	71
5.2 System Design .....	71
5.2.1 Correlation-based Country-Product Matrix .....	73
5.2.2 Anomaly Time Series .....	75
5.2.3 Choropleth View .....	76
5.2.4 Trade Diffusion Graph .....	79
5.2.5 Clustering and Comparison .....	80
5.3 Usage Scenarios .....	85
5.3.1 2011 East Asia Drought and Flood .....	85
5.3.2 Sustainability - Food Insecurity and Political Instability ....	88
5.4 Discussion .....	92

CHAPTER	Page
6 SPATIOTEMPORAL TRADE NETWORK ANALYSIS – A CASE FOR ADDRESSING SPURIOUS CORRELATION .....	94
6.1 Design Requirements .....	95
6.2 Causality Modeling .....	96
6.3 Visual Analytics Components .....	98
6.3.1 Correlation Matrix .....	99
6.3.2 Causal Path View .....	102
6.3.3 Context of Relationship .....	104
6.3.4 Model Tuning .....	106
6.4 Discussion .....	108
7 CONCLUSION .....	110
REFERENCES .....	111

## LIST OF FIGURES

Figure	Page
2.1	Examples of Causal Diagrams. . . . . 11
3.1	The Visual Analytics Framework. . . . . 20
4.1	Semantic Annotation Framework Exploring the Period Between Oct.12 and Nov.2, 2014 of the Climate Change Media Dataset and the ACLED Dataset. The Timeline Shows Annotations of Possible Drivers of Media Reports on Climate Change Framed Around Food Insecurity. Keywords “Agriculture” and “Food” Are Used for the Semantic Mapping Between the Media Topic and the Text Content in The ACLED Dataset. The Causality Model Indicates a Goodness of Fit $R^2 \approx .8$ with $p\text{-value} \leq .05$ . Actors, Locations and Descriptive Text Are Annotated on the Timeline. . . . . 27
4.2	The Topic Keyword View Shows the 50 Most Frequent Topic Keywords for <i>Economy</i> in the Social Unrest Media Collection. Keywords Are Ordered by Their Frequency in (a) and by Topic Significance in (b). Color Refers to the Other Measure Not Ordered by. The Two Bars of Each Keyword Represents Its Frequency Inside and Outside the Topic Respectively. . . . . 30
4.3	The Cluster Force Layout. . . . . 35
4.4	This Figure Shows an Interaction Process of Steering the Cluster Force Layout to Update the Word Concept Clusters. Each Step Is Marked Numerically on the Figure. User Can Drag a Word Away from All Clusters or Towards Another Cluster. User Can Drag a Group of Words to Select Them. . . . . 38



Figure	Page
4.5 Interaction on the Event List View to Update Word Weight and the Event Semantic Similarity. . . . .	39
4.6 Initial Causality Results Without Filtering Events. The Causal Arrows Are Colored Based on Their Significance, and the Current Result Shows No Significant Causality Between the Current Media Stream and Events, and the Best Fit Model Is Displayed with $R^2 \approx 0.6$ and $0.1 < p\text{-value} \leq 0.5$ at a Lag of 4. The Height of the Area Curve Represents the Amount of Variance Explained by the Model. . . . .	41
4.7 Example of the Bipartite View. This View Shows the Events (Left Side) and the Media Articles (Right Side) Indicated by the Causality Lag Under Analysis and Connected by Semantically Matched Keywords. . . . .	46
4.8 Investigating the Plausibility of Climate-Induced Civilian Abuse during the 2014 GHoA Drought. The Link between Violence Events and Social Unrest RSS Have Been Explored Although Casualty Test Shows Non-Significant Result. . . . .	48
4.9 The Annotation Glyph for Event and Media Articles. The Analyst can Annotate Entities by Clicking on the Expanded Nodes. . . . .	49
4.10 Semantic Word Clustering and Filtering for the Study of Climate-Induced Civilian Abuse During the 2014 GHoA Drought. . . . .	54

5.1	A Visual Analytics System for Exploring Global Trade Networks and Their Relationship to Regional Instability. The Anomaly Time Series (1) on Top Displays the Time Series and Anomalies of the Trade Attributes and Stability Measures for a Selected Country. The Choropleth View (2) Along with the Small Multiple Maps (3) Display the First Order Trade Relationships Centering on Selected Countries. The Colored Bars Below Each Small Multiple Map Show the Temporal Correlation of the Selected Trade Measure to the Stability Measure. The Clustering View (4) Displays the Hierarchical Clustering for the Countries, Based on Either the Triadic Similarity or Top Partner Similarity. The Groups Can Also Be Configured to Show the Average Triad Distributions (As (6)) and Other Measures. The Trade Diffusion Graph (5) Displays the Propagation Effect of Anomalies. Connections Between the Nodes Indicate the Import Dependency From The Target Node to the Source Node Is Larger Than a Threshold, Which Can Be Adjusted Using the Slider on the Right. ....	61
5.2	The 13 Possible Triad Configurations Are Labeled with Numbers in Braces. In the First Row, Triad 1 to 6 Are Open Form Triads, While in the Second Row, Triad 7 to 13 Are Closed Form Triads. ....	68
5.3	The Country-Product Matrix view Has an Abstract View (A) and a Detail View (B) to Show the Correlations Between Countries' Trade and Stability Measures. This Figure Shows Cameroon and Mauritania Have Many Correlations Between Cereal Trade Measures and ACLED Events.....	74

- 5.4 The Main Choropleth Map Can Be Used to Visualize Triad, Trade Quantity/Value, and Country Dependency/Leverage. Examples Show the Triad Distribution (Left), China’s Dependency (Middle), and China’s Leverage (Right) in Cereal Network..... 77
- 5.5 The Matrix View Showing the Triad Distribution for Each Country. Rows in the Matrix Represent Triad Configurations and Columns in the Matrix Represent Countries. Each Cell in the Matrix Represents the Triad Count for The Country. The First Column in the Matrix Is the Currently Selected Country in the Map View. The Other Columns Are Ordered Based on Their Vector Similarity to the First Column. The Analyst Can Remove Some Triads (Rows) from the Matrix by Clicking on the Triad on the Similarity Control Panel. Doing So Will Recompute the Vector Similarities and Reorder the Matrix. The Bar Chart in the Bottom Shows the Instability Measure (e.g. ACLED, GDP). 81

- 5.6 The Clustering View Displays the Clusters of Countries in a Dendrogram. Countries in a Cluster Are Grouped in a Box (1). Clicking on Any Box Will Break the Cluster into Two Smaller Clusters. Clicking on Any Internal Node (2) Collapses All Children. The Analyst Can Use the Slider (4) to Adjust the Similarity Threshold for the Hierarchical Clustering. The Color of the Countries Encodes Their Selected Stability Measure. By Clicking on the Setting Icon (3) at the Top Right Corner of Each Box, the Analyst Can Choose Between Different View Options for the Fox. The Box Can Change to a Histogram of the Stability Measure (5) or a Bar Chart That Compares the Mean and the Standard Deviation of Every Triad Configuration (6). . . . . 83
- 5.7 Comparison Between Niger (1) and Tunisia (2) on Each Country's Import Quantity of Rice. By Comparing Both Countries' Major Exporters of Rice Using the Choropleth Views (1a, 2a), It Can Be Seen that Niger Imports from a Wide Variety of Countries Whereas Tunisia Imports Primarily from India and Pakistan. The Time Series for Both Countries (1b, 2b) Show That Tunisia Had an Import Quantity Dip in 2012 While Niger Had No Such Disruption. At the Same Time, Tunisia Had a Sharp Increase in Local Conflict Events. The Analyst Can Also Observe That Tunisia Had Fewer Types of Triads Than Niger. 86

Figure	Page
5.8 Egypt Had a Sharp Increase in Cereal Import Value in 2011. This Figure Shows (1) the Import Value Distribution of Egypt, (2) The Similar Countries Who Had the Same Increase, (3) the Scatter Plot by ACLED and Import Value of Cereal Among Africa Countries, and (4) the Time Series Indicating the Sharp Increase as Well as the Negative Correlation Between Egypt's Import Value and Its Stability Index. . . .	89
6.1 The System Workflow. . . . .	98
6.2 The Correlation Matrix. . . . .	100
6.3 The Causal Path View. . . . .	102
6.4 Trend Comparison. . . . .	105
6.5 Variable Selection. . . . .	106

## Chapter 1

### INTRODUCTION

When extreme drought struck China's main wheat-growing regions in 2011, it was not China that suffered food shortages, but far-off countries like Egypt, which experienced widespread civil unrest and food riots [1]. How was this possible? Why was this possible? Why did policy makers not foresee this potential risk to social stability? These questions are all related to finding the co-occurrence and cause-effect relationships between the countries, the international policies, or other underlying features. Causality and correlation analysis have a wide range of applications, such as finding the relationship between a person's education level and their health [2], finding associations between pet ownership and health promotions [3], or identifying if a company's marketing campaign increases their product sales [4]. Sometimes, it is necessary to focus on finding the cause-effect relationships, i.e. causality analysis, because causal relationships theoretically remain stable under any condition [5]. This property can become very important for safety-critical problems. For example, it is important for the drug companies to understand the actual effect of their drug regardless of the conditions of the patient [6–8].

Unfortunately, causality analysis often remains relatively overlooked in the data science community. One reason is that causality detection is still an extremely difficult problem. Causal relations, in principle, cannot be tested solely based on mathematical models [9, 10], and, ideally, controlled experiments need to be carried out to reliably establish causality [11, 12]. However, the cost of these experiments can sometimes hinder their applicability. Thus, researchers often resort to observational studies in which causal relations are extracted from observational data. Unfortunately, models

trained on observational data are susceptible to bias, which makes them much less reliable than those derived from experiments. One way to mitigate this problem is to use expert knowledge to justify the causal claims, which often rests upon some theoretical backgrounds, and these causal relations have to be viewed through the lens of domain specific theories. However, as the data becomes more complex, the utility of causality analysis with this approach tends to drift towards an exploratory nature, in which case visualizations often become desirable supplements.

While causality analysis on observational data has recently received more attention, handling large observational data has always been hard. Problems, such as high dimensionality or a lack of well defined structure, make many datasets difficult to analyze. Furthermore, the daunting complexity of many datasets makes it hard to determine what to explore. In the case of the global food trade network [13], the data contains a trade network with over 200 countries and more than 600 food products. If the analyst's task is to find out whether the decrease of exports of any combinations of trade products from any countries has any effect on the social stability of another country, they would have to examine the relationships between a large amount of pairs of variables. Such tasks can be automated, but in order to identify causality, human input is still required. In another example, when analysts want to analyze a large amount of news to identify whether exposure to certain information has the potential to spur political uprisings, it is common to use natural language processing (NLP) and information retrieval techniques to pre-process the data before feeding it into any statistical models. In this case, the analysts have to make sure the pre-processed data accurately represents the information in the text, otherwise any correlation derived from the data could be spurious. It is very likely that some human intervention is needed to fine tune the pre-processed results. The size and complexity of these observational data can also intensify the deficiency of the causal models,

whose quality and applicability tends to degrade significantly with the number of variables. As the number of variables increase, these models take longer to execute and become much harder to interpret. Thus, some feature selection processes are required for pre-process, ideally with the intervention of human experts, as machines do not yet fully possess the power of human intuition.

In order to develop an effective approach to assist people in hypothesis generation and correlation and causality analysis, this thesis develops a visual analytics framework for causality analysis. Visual analytics serves as a means of bridging automatic models and the end users to interactively support exploring the data, model, and results. In causality analysis, exploration and interaction can serve as a way to inject user knowledge into the analysis process to help identify the “true causality”. This framework facilitates causality analysis by developing several interactive visual analytics components to assist exploration, variable extraction, and model tuning. Case studies of this framework were conducted to demonstrate its capabilities across different domain applications.

The first case study focuses on the exploration, linkage, and annotation of multiple media sources to explore drivers of discourse. In this case study, text analysis methods were used to extract relevant media documents and the results were subsequently used for causality analysis. First, semantic matching was applied to identify keywords and concepts that an analyst considers to be related between two datasets. A novel widget enables domain experts to quickly cluster, split, and merge keywords from a semantic dictionary to ensure that meaningful similarities are captured through a visual to parametric interface while allowing for analyst-guided language disambiguation. While there are known limitations of keyword searches, by enriching an analyst’s choice of keywords with semantic meaning, a broader matching was enabled that better aligns with the user’s mental model. In this way, searching was not lim-



ited on one (or several) keyword(s) but instead applied on semantic meanings that embeds the analyst’s domain knowledge. Once users are satisfied with the semantic grouping of keywords, filtering is performed and a raw count of semantically related articles and events per time step can be extracted from each of the time-oriented textual datasets. Using these time series, a secondary annotation step is performed where causality measures are applied to the derived time series to extract possible drivers. These causality measures could indicate that past events contained in time series  $A$  contain information that can help predict time series  $B$  above and beyond the information contained only in time series  $B$ . If a causal link is established, the framework then indicates the temporal lag under which causality was identified and provides interactions to further filter and annotate the time series based on relationships between locations, actors and other derived information.

The second case study focuses on enabling the users to understand how trade relationships impact local vulnerabilities over time in a global trade dataset. To support such an analysis of multivariate trade data, the visual analytics framework integrated a local network structure analysis technique based on triadic closure [14], and anomaly detection and correlation analysis for pattern analysis and hypothesis generation. To provide a basic understanding of the trade data, an interactive global map view was used to visualize both the volume and proportion of trade events for imports, exports, and triadic structures. Temporal correlation analysis and anomaly detection are provided to guide the user to structures of interest within the data, and small multiples are used to allow comparisons between regions with similar data features. Other views include a triad comparison matrix which compares the triad distribution for each country, and a triad hierarchical clustering view which groups countries based on their triad profile similarities.

The third case study extends the trade analysis case and focuses on identifying spurious correlations. Here, causal models were trained on trade measures and stability measures, and the results were matched against the detected correlations to identify potential spuriousness. The corresponding sub-network of the causal model was visualized to help the analysts understand the rationale of the indicated spuriousness. Additional contextual information surrounding the causal relations was provided to help inspect the causal relations, and interactions were implemented to adjust the models.

In conclusion, this thesis develops a visual analytics framework for correlation and causality analysis which assists analysts in exploring and manipulating data and models. This framework structures a general visual analytics pipeline for causality analysis, and specifies the role of interaction between human and machines in correlation and causality analysis. Three case studies were conducted based on this framework, and demonstrate that the visual analytics framework is effective at assisting the analysis. The contributions of this thesis include:

1. A general framework for visual analytics in causality analysis.
2. Applied causality metrics for identifying topic drivers in media streams.
3. A user-guided semantic lexical matching scheme for document selection.
4. A causality-driven annotation scheme for exploring potential media drivers.
5. A novel triad analysis for exploring network dependencies to identify possible relationships between trade network structures and other measures of interest.
6. An integration of space, time, and network analysis methods to support the linked analysis of trade network data and conflict events through anomaly detection and correlation analysis across imports, exports and triadic structures.

7. An end-to-end solution to causality analysis, which makes causality analysis more accessible by facilitating the generation, exploration, and modification of causal models.

### BACKGROUND AND RELATED WORK

Causality is a basic metaphysical notion concerning the relationship between cause and effect [15]. Philosophical thinking about causality predates those of Aristotle, who extrapolates that causality analysis is the search for explanation about states of being [16]. Modern theories on causality, however, were developed in the past century, which progressed through a diverging path led by Rubin and Pearl. Rubin [11] pioneered the usage of the potential outcome framework [17] for causality analysis in which causality is expressed as the difference between two potential outcomes, with one of them being the observation. The effect of the potential outcomes are typically estimated through randomized experiments. Pearl laid the foundation of graphical causal models [5] in which causal relations are expressed as conditionally independent relations that can be fully described in a directed acyclic graph (DAG), often referred to as the Bayesian network [18]. Casting all causal assumptions in a graphical language offers a more complete and intuitive representation of causal dynamics among variables, and it also makes it easier to algorithmically identify causal effects based on non-experimental data [5, 19]. Pearl’s work led to a proliferation of causality analysis techniques for observational data [20–22], which spans a wide range of adaptations in fields such as epidemiology [23–27], biology [28–30], and social science [26, 31]. Both the potential outcome framework and graphical causal model address problems regarding the search for “true causality”, whereas some works instead focuses on generating causal hypothesis. One such line of work emerged from econometrics, led by Granger with his Granger causality method [32]. Granger causality is primarily concerned with the predictive ability of one time series over the other, and,

despite its name, is really not about “true” causality due to its inability to control for unobserved spurious effects. However, by providing evidence regarding temporal precedence, Granger causality is a powerful method for hypothesis generation, and its simplicity and interpretability have contributed to Granger causality’s popularity in temporal causality analysis. Recently, there have also been efforts to combine Granger causality with graphical models to address its susceptibility to spurious effects [33, 34]. This thesis leverages Bayesian networks and Granger causality for causality modelling, combining them with visual analytics components to support causal explanations and hypothesis generation.

Visual analytics systems dedicated to causality analysis are still in their infancy, but there has been a rich body of visual analytics research focusing on correlation analysis. Correlation analysis can serve as a preliminary step for causality analysis and can also provide insights for data exploration. This thesis also uses correlation analysis as means of insight generation, and visual analytics components are developed for this task. The following sections will discuss correlation visualizations, graphical models on causality analysis, and the current state of the art in causality visualization.

## 2.1 Correlation Visualizations

Causality analysis can begin with identifying correlations, and correlations can provide many hints about the underlying patterns of the data. Correlation analysis and visualization are essential components of this thesis and significant effort by analysts are often spent on correlation inspection on high dimensional spatiotemporal data. Many visualization techniques have been established for correlation analysis, such as the scatterplot, the parallel coordinate plot [35] and the correlation matrix. While these visualizations provide users with an overview for simple tasks, more complex frameworks are needed for in-depth analysis. Zhang et al. proposed

a visualization framework called Correlation Map [36] to visualize correlations for multi-variate data. Correlations among variables were visualized in a network using a force-directed layout with the strength of force encoding the strength of the correlations. A hierarchy of variables was also built using a correlation-related metric to enable splitting and merging operations on the variables. The authors also extended the interface to include a subspace scatter plot to integrate data into the Correlation Map [37], which projects the data points to the divided triangles from the network. Qu et al. [38] introduces a polar system using a circular pixel bar chart to detect correlations between wind direction, wind speed and other attributes to help users understand how these facts contribute to the air pollution problem in Hong Kong. Correlation based systems can also be used for inspecting high dimensional data. Xia et al. [39] developed a visual analytics system that explores the relationships between dimensions by comparing the similarity of views, which serves as the basis for feature selection and categorization. Yuan et al. [40] visualized correlations of dimensions by projecting them into a 2D plane, where these dimensions are first grouped in a hierarchy, and each group can be subsequently compared. Visualization has also been used for inspecting temporal correlations. Dang et al. [41] developed a system that visualizes correlation among dimensions in a high dimensional temporal dataset, in which temporal comparisons are done through a sequence of scatterplots and dissimilarity measures of these scatterplots are computed and visualized in a time series. Lee et al. [42] designed an algorithm to extract and cluster salient temporal trends and visualizations were developed to display these temporal trends. The methods proposed in these works are not directly used by this thesis, but correlations on large spatiotemporal data are explored in this thesis, and this thesis will discuss its own approaches to correlation visualization.

## 2.2 Causality Analysis

It is well known that “correlation does not imply causation”. However, the majority of correlation visualization tools does not address this fundamental issue in causality analysis. That is, how can we make the analysts to determine if the detected correlation is spurious. For this, there needs to be some addition to correlation visualization that stimulates causal reasoning. This thesis attempts to incorporate graphical causal models as a guidance to causal reasoning in the proposed visual analytics framework. This section provides general background on graphical causal models, along with some techniques that can be used to generate and manipulate these models, many of which are related to the methods used in this thesis.

### 2.2.1 Probabilistic Causation

Causality analysis, as opposed to correlation analysis, attempts to find fundamentally stable directional relationships that would hold true regardless of the conditions binding them [10]. In order to verify a causal relationship, one would have to examine all factors that can potentially influence the causal relations, which leads to the philosophical notion that true causal relationships can not unequivocally be found. However, by limiting the problem to a restricted setting, one would be able to assess the probability of something causing the other [43].

One way to express causality probabilistically is to use conditional probability [44–46]:

$$X \rightarrow Y \iff P(Y | X) > P(Y) \tag{2.1}$$

In other words,  $X$  causes  $Y$  if and only if the occurrence of  $Y$  given  $X$  is more likely than that of not given  $X$ . However, this definition suffers from two fundamental deficiencies [47]: the equivalence of  $P(Y | X) > P(Y)$  and  $P(X | Y) > P(X)$  means

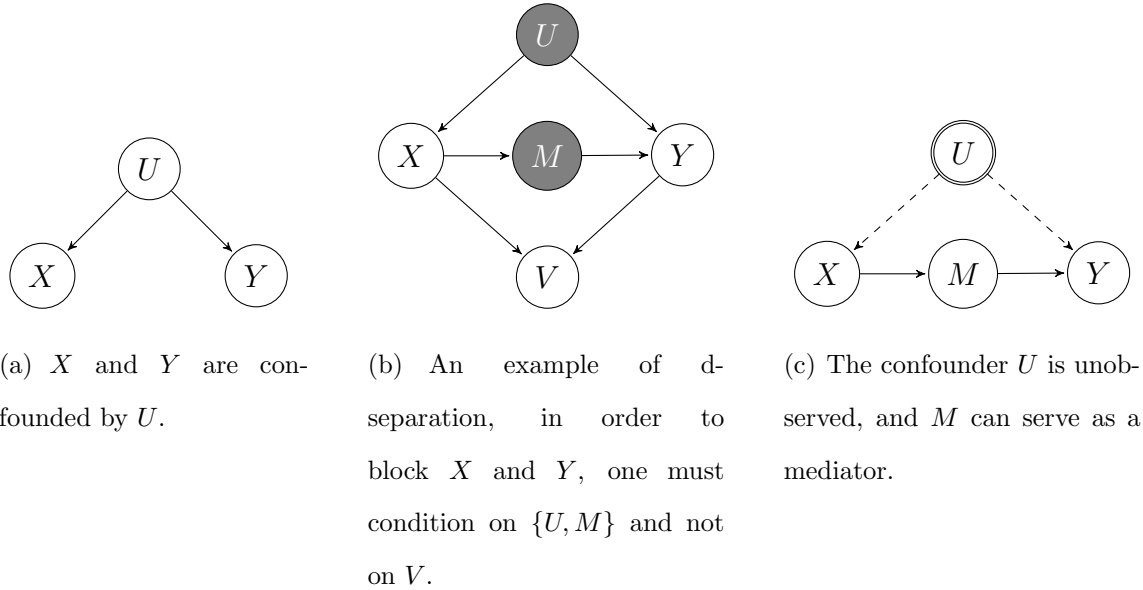


Figure 2.1: Examples of Causal Diagrams.

the directionality of the relationship can not be deduced from Equation (2.1) alone; and the potential spuriousness of the relationship has not been adequately addressed. In other words, Equation (2.1) is essentially an expression of association, thus more restrictions are likely needed in order to specify causality.

One way that spurious correlations between  $X$  and  $Y$  can arise is due to variables called *confounders* [5, 48], which are defined as factors that have influence on both  $X$  and  $Y$ . These confounders can introduce a magnifying influence, namely *confounding bias* [5], on the relationship between the pair of variables. In an extreme case, shown in Figure 2.1(a), the confounder  $U$  can make two completely unrelated variables  $X$  and  $Y$  appear positively correlated. In order to prove/disprove the existence of a causal relationship, one must condition on the confounders to remove their effects. In Figure 2.1(a),  $X$  and  $Y$  are said to be conditionally independent of  $U$ , expressed as [5]:

$$P(Y | U) = P(Y | X, U)$$



or

$$(X \perp\!\!\!\perp Y \mid U)$$

Therefore, the correlation will completely disappear if one conditions on  $U$ . However, it is often the case that the causal effects coexist with the effect of the confounders, in this case, conditioning will expose the actual causal relationship by removing the confounding bias.

Another more peculiar cause of spurious correlation results from having  $X$  and  $Y$  influencing the same external factor(s)  $V$ , namely the *collider* [10], in which case spurious correlation between  $X$  and  $Y$  arise when one incidentally conditions on  $V$  during analysis. This phenomenon is also called Berkson's paradox [49].

Conditional independence can be derived using a graphical criterion introduced by Pearl [5, 18] called *d-separation*. Given a set of relationships as a graph,

*“A path  $p$  is said to be d-separated (or blocked) by a set of nodes  $Z$  iff*

- 1.  $p$  contains a chain  $i \rightarrow m \rightarrow j$  or a fork  $i \leftarrow m \rightarrow j$  such that the middle node  $m \in Z$ , or*
- 2.  $p$  contains an inverted fork (or collider)  $i \rightarrow m \leftarrow j$  such that the middle node  $m \notin Z$  and no descendant of  $m$  is in  $Z$  [5].”*

An example of this concept is illustrated in Figure 2.1(b), in which case  $X$  and  $Y$  are d-separated by the set  $\{U, M\}$ , which is equivalent to saying  $(X \perp\!\!\!\perp Y \mid UM)$ . It is also worth noting that d-separation involves the notion of *chain*, such as  $X \rightarrow M \rightarrow Y$  in Figure 2.1(b). The idea that  $M$  intercepts  $X$  and  $Y$  stems from the fact that by knowing the state of  $M$ , the information on the state of  $X$  would become redundant for assessing the state of  $Y$ . However, one can also stress that  $X$  still asserts influence on  $Y$ , albeit indirectly. When the analysts are solely concerned about assessing the

actual influence from one variable to the other (e.g.  $X$  to  $Y$  in Figure 2.1(b)) — as is for the majority of research, much attention should be directed to controlling the confounding bias. It is for this reason that Pearl [5, 50] introduced an extended concept called the *back-door criterion*. In a directed acyclic relational graph (DAG),

“A set of variables  $Z$  satisfies the back-door criterion relative to an ordered pair of variables  $(x_i, x_j)$  if:

1. no node in  $Z$  is a descendent of  $x_i$ ; and
2.  $Z$   $d$ -separates every path between  $x_i$  and  $x_j$  that contains an arrow in to  $x_i$  [5].”

As for the example in Figure 2.1(b), the set  $\{U\}$  is said to satisfy the back-door criterion relative to  $\{(X, M), (M, Y)\}$ . This allows us to assess the effect of  $X$  on  $Y$  with the equation:

$$P(Y | X) = P(Y | X, U)P(U)$$

With each component  $P(y | x)$  computed as a marginalization over  $U$ ,

$$P(y | x) = \sum_u P(y | x, u)P(u)$$

However, it is often the case that the confounder(s) are unobserved and/or their effects are very hard to measure [51], which makes it nearly impossible to directly assess the causal effect between a pair of variables. Fortunately, the effect of causal relations can still be estimated with observed data through the help of intermediate variables called *mediators* [52–55]. As an illustration, consider modifying the example in Figure 2.1(b) such that the variable  $U$  represents a set of latent confounding factors influencing both  $X$  and  $Y$  (Figure 2.1(c)). Then, with the assumption that  $X \rightarrow M$  is unconfounded and  $M \rightarrow Y$  is unconfounded given  $X$ , we can leverage the mediator  $M$

to directly estimate the effect of  $X$  on  $Y$ , circumventing the latent confounder(s)  $U$ . Pearl [5, 56] summarized the requirements for mediators in his *front-door criterion*.

“A set of variables  $Z$  is said to satisfy the front-door criterion relative to an ordered pair of variables  $(X, Y)$  if:

1.  $Z$  intercepts all directed paths from  $X$  to  $Y$ ;
2. there is no unblocked back-door path from  $X$  to  $Z$ ; and
3. all back-door paths from  $Z$  to  $Y$  are blocked by  $X$  [5].”

Given that  $\{M\}$  satisfies the front-door criterion relative to  $\{(X, Y)\}$ , the effect of  $X$  on  $Y$  can be computed as [5]

$$P(y | x) = \sum_m P(m | x) \sum_x P(y | x, m)P(x)$$

### 2.2.2 Bayesian Networks

Pearl’s back-door and front-door criterion theorized the general approaches for controlling confounding bias. A large body of research has implicitly or explicitly included these approaches [57–59], and many of these approaches involve a standard method called the *controlled experiment*, in which the causal relationship is established through measurements on randomized samples with the confounding variables carefully vetted and controlled. To carry out this type of experiment, one would have to assume a theoretical structure of the relationships, then carefully verify them based on experimental results. Obviously, this type of analysis can encounter limitations when dealing with complex multivariate problems as the cost of the experiments can become prohibitively high. Hence, we are seeing an increasing amount of studies that attempt to extract causal relations from observational data, where the results can serve as aids for policy making or guidance for experiments [60–62]. Many of these

studies involve a type of graphical model called the *Bayesian network* [63]. Bayesian networks represent conditional dependence/independence among variables in a directed acyclic graph (DAG), which can be interpreted as a causal networks [5, 63]. Learning Bayesian networks from data generally involves two steps [21]:

### 1. **Structural Learning**

Structural learning is a process to estimate the DAG structure of a network from data, where the methods often build on top of the structural causal theories previously mentioned. There are, in general, two approaches to structural learning [21]: the score based approach and the constraint based approach. The score based approach searches through the space of available structures given the data and finds the one which produces the optimal score, which is typically calculated by solving the maximum likelihood estimation (MLE) for the structure given the data [21]. Since the number of available structures grows exponentially with the number of variables in the data, the search is an NP-hard problem [64]. Some heuristic based methods can also be applied to reduce the search space, such as the hill climb algorithm [20]. These methods often settle in a local maximum, thus their results can only be treated as an approximation.

The constraint based approach, pioneered by Verma and Pearl in their causal discovery framework [65, 66], attempts to find the causal structure by first testing conditional independence between each pair of variables given all possible combinations of parent nodes, which effectively produces an undirected graph called the *skeleton*. Then, the DAG can be obtained by orienting the edges according to the constraints imposed by the skeleton. The conditional independence can be tested using methods such as the  $\chi^2$  test [67] (for discrete variables) and partial correlation [68] (for continuous variables).

## 2. Parameter Estimation

Parameter estimation is a process to estimate the conditional probability distributions (CPDs) given the structure and observation. Each CPD specifies the probability distribution of a variable given its parents in the network, and it can typically be derived via MLE by assuming that the prior probability follows a Dirichlet distribution [69]. A Bayesian network is fully specified if the CPD of each variable in the network is filled.

The learned Bayesian network provides a best guess of the causal structure given the observation, which can serve as a guidance to causality inspection. The utility of this model in the proposed visual analytics framework will be discussed in Chapter 6.

### 2.2.3 *Visual Analytics on Causality Analysis*

There are several works that focused on visualizing causal diagrams. Elmqvist et al. proposed a visualization technique called growing squares [70]. This visualization design borrows from the Hasse Diagram [71], with the exception that each node is assigned a unique color, and an animated growing square is triggered when a causal effect initiates. The growing squares kept a mixture of colored grids representing the cause and effect nodes to keep track of the event sequence. One problem with this layout is that it uses a simple color coding scheme that does not scale well with system size. Elmqvist later proposed an enhanced version called growing polygon [72] where each node is assigned an n-sided polygon, and each such polygon is subsequently filled with the colors of the causes influencing it. However, both growing squares and growing polygons have limited abilities for signifying causal strength. Kadaba et al. [73] address these problems by depicting causal relations by node-link arrow and glyphs, leveraging simple animations of node sizes to indicate interactions between the

factors and the target. Other work has been done to directly visualize the Bayesian belief networks [74] in which the layout is guided by a temporal order, and multiple visual variables like color, node size, and proximity are used to represent network semantics.

Causality visualization can also be used as an aid for causal inferences. Wang et al. [75] developed a visual analytics system utilizing Bayesian networks to help incorporate human knowledge in identifying causal relationships. In their system, the causal relationships were visualized using a force directed layout [76]. The analyst is allowed to interact with the causal graph interface to modify the diagram, such as adding, removing, or modifying a causal relationship. Visual feedback is provided to help the analyst assess the implications of such modifications. However, the drawback of this layout is that it can potentially display an overly complex structure, especially when the diagram is large. Following this, Wang et al. proposed an improved version [77] which contains a layout that emphasized the flow of causal sequences. This layout extracts causal paths from the diagram using spanning trees and displays them as a flow in consistent directions. In addition, the system also allows for comparison of models trained on subdivisions of the data, along with a visual pooling process to summarize and diagnose different models. One limitation of this approach is that it cannot analyze causalities in time series data. For time series data, Granger Causality [32] is a widely adapted technique. Zhu et al. [78] proposed a spatio-temporal Granger Causality model to analyze causality between urban dynamics and air pollution. However, there have been few works utilizing Granger Causality in the visual analytics domain. This thesis utilizes both Bayesian networks and the Granger causality for causal modelling and visualizations are implemented to assist the analysts understanding and manipulating the models. Details of the implementations will be discussed in the case studies, and they differ from Wang and Zhu's

work by providing contextual information surrounding the detected causal relations to further enhance understanding of the causal relations.

### VISUAL ANALYTICS FRAMEWORK FOR CAUSALITY ANALYSIS

Causality analysis with observational data has become more popular due to increasing availability of data, but its applicability is still limited due to several constraints:

- Causal models trained on observational data are susceptible to bias due to spurious effects from unobserved variables. This unreliability has put people off from using these models in critical tasks. However, with enough expert oversight, these causal models can turn into powerful tools to solve problems that were previously deemed impossible. Unfortunately, to the best of my knowledge, there do not exist many tools to assist human experts in inspecting and interpreting the model results in an interactive and intuitive way.
- Learning causal models with large amounts of data can be very time consuming, and the interpretability of the model tends to decrease as the number of variables gets larger. In order to address this issue, one would have to omit certain variables that do not contribute to the analysis to reduce the complexity of the model. This process cannot be completely automatized, as it will put further strain on interpretability.

The goal of the proposed visual analytics framework is to assist causality analysis by easing the interactions between human experts and automatic models, and to help the human experts better understand and manipulate the data and model results. Doing so can potentially mitigate the issues of interpretability and make causality



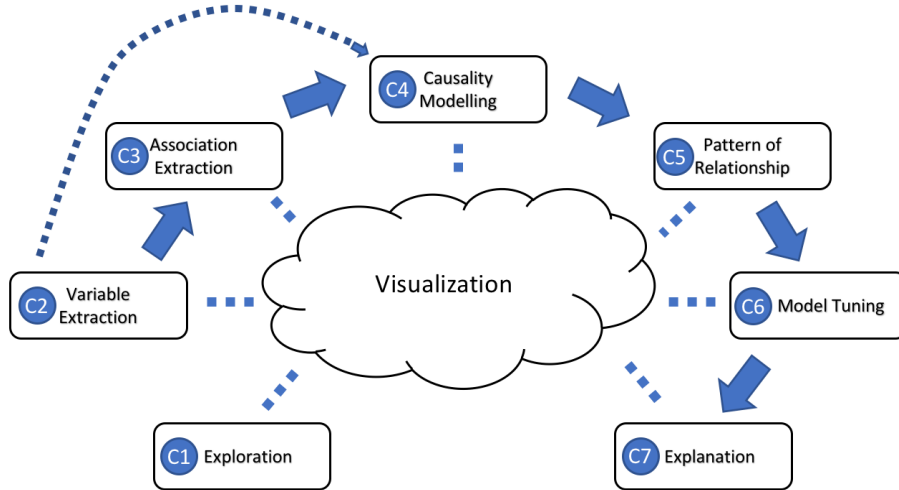


Figure 3.1: The Visual Analytics Framework.

analysis more accessible. In order to achieve this goal, the framework needs to satisfy several design requirements:

- R1** The framework needs to support exploration and manipulation of data to improve insight generation and modelling.
- R2** The framework should enable interactive training and tuning of models and support immediate feedback.
- R3** The framework should provide contextual information surrounding model results to assist explanation and verification of the results.

Figure 3.1 illustrates the pipeline of the framework, which consists of the following components:

### C1 Exploration

The typical analysis workflow begins with exploring the dataset. Analysts rely on this approach to develop an understanding of the data and perform qualitative measurements to help generate insights to guide the direction of the

analysis. Many problems in the dataset can also be discovered during the exploration, such as missing or erroneous data. Exploration is not only limited to the initial stage of analysis, but exploration can help enhance the understanding of the model output throughout the analysis. Visualization has played an important role in exploratory analysis because it is effective at summarizing and presenting large volumes of data [79, 80]. This framework uses visualization to facilitate exploration on data and models, as well as the intermediate results from each component in the pipeline.

## **C2 Variable Extraction**

Variable extraction is a process to refine and extract useful features from a dataset to be fed into the analytical process. This process typically involves filtering redundant features and/or transforming the values of the data to produce a concise and accurate representation of the data, it also often involves organizing and standardizing the data format when the data is unstructured. Variable extraction is a necessary step for nearly all types of numerical analysis; however, there exists some difference in preference when it is applied in causality analysis versus predictive analysis. In predictive analysis, the goal is to predict future observations, thus variable extraction tends to focus on selecting the subset of data that helps produce the most optimal predictions, sometimes by sacrificing the interpretability of the model. In causality analysis, the focus of variable extraction is to select the set of variables that mirrors the participating entities in the phenomenon being studied in order to produce a comprehensive model that best explains the phenomenon, in which case sacrificing interpretability is much less acceptable. In causality analysis, variable extraction can be supported by incorporating domain knowledge, which makes visual analytics very applicable.

### **C3 Association Extraction**

Association extraction [81] can be defined as a process to model the relationships between the variables of interest, regardless of if they are causal or not. Association extraction can serve as a preliminary step for causality analysis, but association extraction itself can also provide valuable insight. Common methods for association extraction include correlation analysis and regression. However, the analysts need to be aware that some of the correlations can be spurious and additional examinations may be needed to rule out these spurious correlations. To carry out these examinations, the analysts can either rely on their domain knowledge and/or conduct experiments. However, as the data size gets larger, it becomes increasingly difficult to manually examine each pair of correlations, thus some automatic methods are necessary.

### **C4 Causality Modelling**

Causality modeling attempts to model the true causal relationships among variables. This process often involves two steps: separating the true relationships from the spurious ones (as mentioned in the previous section) and determining the direction of the relationship. The most reliable way to identify true causal relationships is through controlled experiments. However, as the problem gets more complex, the cost of running these experiments tends to grow as well. Recently, there has been a growing interest in running causal models on observational data, commonly by using Bayesian networks and/or structural equation models (SEM) [23–27]. These models attempt to capture causal relationships by dissecting the structure of the relationships among the variables. However, the results they generate are not guaranteed to be true causality due to spurious effects imposed by externalities beyond the observations. Moreover,

the direction of causal relationships generated by these models can also be ambiguous since there could be multiple “optimal” orientations that fit the same observation. Because of these shortcomings, it is extremely important to double check and verify the results of these models, which often requires human input.

## **C5 Pattern of Relationship**

Since causal models trained on observational data are inherently ambiguous, it is up to the analyst to decode the meaning of the causal relationships and decide whether or not they are reasonable. As such, it is necessary to examine these relationships through the lens of a broader context. Although the analysts can often supplement those information based on their own domain knowledge, it is still desirable to provide contextual information in the system targeting the causal relationships. This is not just for convenience, since large amounts of data can be collected through existing infrastructure, the analysts can clearly leverage these data to gain a better understanding of the relationships and discover potential surprises.

## **C6 Model Tuning**

If there are invalid causal relationships in the model, it makes sense to allow the analyst to modify the model. As such, the framework should support operations such as adding/removing variables, adding/removing causal relations, and setting variables to constants. Doing so may alter the optimal structure of the relationships between the variables, which means the causal relations in the model will need to be adjusted accordingly. The analysts then have to investigate the validity of the updated model by going through the pipeline again.

## C7 Explanation

Once the analysts are satisfied with the model, they can then try to explain the model with their own knowledge and/or the information provided by the system. The system can also include functionality to allow the experts to annotate the final presentation of the results. These annotations can help experts immediately interact with the data to flag events that can constitute changes in the underlying equilibrium of these processes, they can also potentially help the experts to save the states of their “thinking process” on the findings so they can share these with other experts to ease communication and collaboration.

The framework in Figure 3.1 is a general visualization pipeline for causality and correlation analysis. The following chapters will discuss three case studies that implements visual analytics systems based on the proposed framework to address to specific problems. One case study focuses on identifying causal relationships between conflict events and media reports by extracting information from two text documents. The other case studies focus on exploring correlation and causality between trade and social stability in every country from a large spatiotemporal food trade network data.

## IDENTIFYING TOPIC DRIVERS IN MEDIA EVENTS

As citizen news reports, micro-blogs and other media outlets have increased, a variety of tools have emerged for analyzing media data collections. Such tools tend to focus on topic extraction [82, 83], event detection [84], and information flows [85, 86] as a means of quickly assessing the development of ongoing stories. However, recent work often focuses on exploring evolving media discourse in isolation. What is needed are tools and methods that can enable analysts to link together multiple data sources of interest in order to identify patterns and drivers that exist between datasets that are not fully captured or represented in any single dataset alone. To that end, new technologies have been developed for fusing media data and secondary data sources to provide contextual information. For example, recent work from Wanner et al. [87] and Hullman et al. [88] explored methods for time series analysis to identify text features of interest in conjunction with quantitative phenomena observed in stock prices, Liu et al. [89] proposed *TextPioneer* with a combination of hierarchical tree visualization and a twisted-ladder-like visualization to present and analyze the lead-lag patterns in a topic between different corpus, and work by Lu et al. [90] explored methods for identifying intervention points in news media data to cue analysts into the exploration of secondary datasets of interest.

However, fusing datasets and providing means of identifying and annotating potential temporal drivers is still fraught with challenges. For example, imagine collecting a corpus of text from Twitter discussing a sale product (e.g., tennis shoes) as well as a collection of product reviews on tennis shoes from Amazon. In this case, an analyst may want to see if discussion on Twitter is driving ratings and comments

in the product reviews. Challenges here could be that the language used on Twitter and the language used on the product review site do not have a one to one matching (e.g., “This shoe is sick” could be counted as a positive review, but the language on the product review site may use less slang). As such, keyword searches to filter the document collections to only positive reviews may not be able to rely on traditional topic modeling tools and often need domain expert intervention. Once a dataset is curated, then further automated analysis to explore drivers between the datasets must be performed (for example, do positive tweets about a product proceed positive product reviews?). Annotations of key events in the timeline and key actors in the text corpora are also relevant and need to be annotated in the hypothesis exploration and analysis phase to help the analyst navigate large document collections and identify key components of the event drivers.

As such, this case study instantiates the proposed visual analytics framework (Figure 3.1) into a system (Figure 4.1) that focuses on the exploration, linkage, and annotation of multiple media sources to explore drivers of discourse. First, semantic matching was applied to identify keywords and concepts that an analyst considers to be related between two datasets. A novel widget enables domain experts to quickly cluster, split, and merge keywords from a semantic dictionary to ensure that meaningful similarities are captured through a visual to parametric interface while allowing for analyst-guided language disambiguation. While there are known limitations of keyword searches, by enriching an analyst’s choice of keywords with semantic meaning, a broader matching was enabled that better aligns with the user’s mental model. In this way, the system moves away from searching on one (or several) keyword(s) and

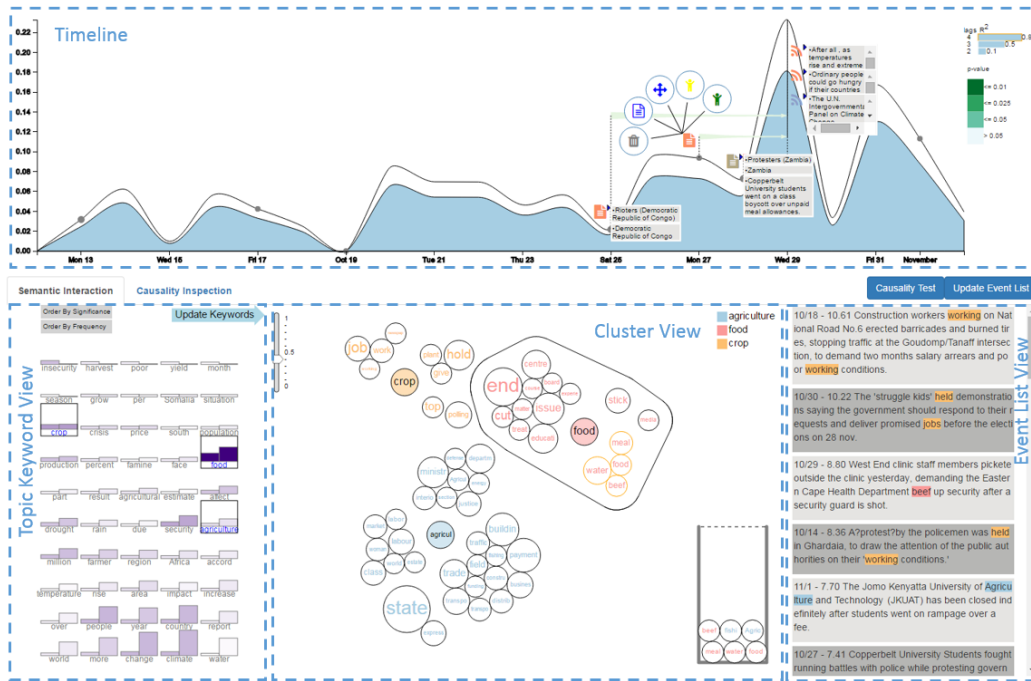


Figure 4.1: Semantic Annotation Framework Exploring the Period Between Oct.12 and Nov.2, 2014 of the Climate Change Media Dataset and the ACLED Dataset. The Timeline Shows Annotations of Possible Drivers of Media Reports on Climate Change Framed Around Food Insecurity. Keywords “Agriculture” and “Food” Are Used for the Semantic Mapping Between the Media Topic and the Text Content in The ACLED Dataset. The Causality Model Indicates a Goodness of Fit  $R^2 \approx .8$  with  $p\text{-value} \leq .05$ . Actors, Locations and Descriptive Text Are Annotated on the Timeline.



instead search on semantic meanings that embeds the analyst’s domain knowledge into the search.

#### 4.1 Analytic Tasks

As an example analysis, first consider a scenario in which a political scientist wants to explore the relationship between local conflicts and the 2015 Nigerian election. Prior to the Nigerian election, domain experts had hypothesized that widespread riots and social unrest would occur as part of the election cycle. The domain expert wanted to analyze social unrest news articles and local conflict events in Africa during the time leading up to the election to inspect if the election campaign media was a precursor to violence (or vice-versa). First, the data for analysis is collected from two sources (news media and a curated event dataset documenting violent conflicts in Africa). These datasets are a superset of the data needed for analysis. Thus, the first task is filtering the data for the relevant text. Once a subset of the media posts and the conflict event records are curated, the documents need to be checked for relevancy, and a temporal aggregation must be performed to enable cause-effect relationship analysis. In this case, the analyst is interested in relationships between the subset of media data related to the election in Nigeria and the subset of conflict events in Africa occurring in Nigeria. As key events, actors and locations are identified in the data, the analyst also needs to annotate their findings in order to explain the events and the news contents and support the hypothesis generation and explanation phase.

Given this example analysis task, several components of the proposed visual analytics framework (Chapter 3) can be used to create a work flow that can be generalized into several key steps for the identification of topic drivers in media data:

1. Identify topics within the datasets in order to explore their evolution over time (Ch. 3: **C1**);
2. Link and filter datasets so that the extracted media items are relevant to the analysis (Ch. 3: **C2**);
3. Identify critical entities within the datasets (Ch. 3: **C2**), and;
4. Provide methods for cause-effect identification and identify potential leading or lagging indicators (Ch. 3: **C3, C4**).

## 4.2 Semantic Event Retrieval

### 4.2.1 *Semantic Lexical Match*

In many application areas, the key to successful data analysis and reasoning involves integrating data from different sources, for example, linking financial data with news reports may help analysts develop models for predicting stock market responses [87]. One critical task in linking multiple datasets is performing text query matches based on document similarities. This task usually leverages information retrieval methods [91]; however, many media posts contain short text messages or other limited information which may not contain the specific query word, restricting the effectiveness of simple keyword matching. In order to solve this problem, word semantic similarity measures have been studied [92, 93]. The general idea is to match the word sets from text segments by pairing every word in one data collection with its most similar word from the other data collection and then calculate a weighted sum of all pairs. Though this method can be used to measure the semantic similarity between text segments, there are two major challenges in querying for relevant events in media:

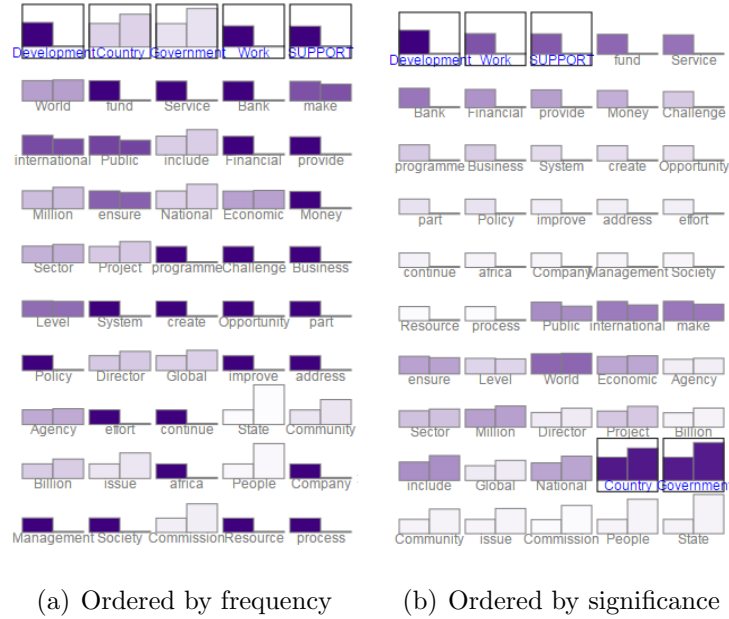


Figure 4.2: The Topic Keyword View Shows the 50 Most Frequent Topic Keywords for *Economy* in the Social Unrest Media Collection. Keywords Are Ordered by Their Frequency in (a) and by Topic Significance in (b). Color Refers to the Other Measure Not Ordered by. The Two Bars of Each Keyword Represents Its Frequency Inside and Outside the Topic Respectively.

1. One dataset under exploration may have a wide word coverage which would tend to increase the maximum similarity measurements between the two datasets.
2. The knowledge-based word similarity measure returns the highest similarity score among all possible word senses [94]; however, not all word senses are relevant to the analyst’s semantic definition. For example, the word “crop” can relate to agriculture or a type of haircut, but an analyst studying agriculture would likely be uninterested in articles about hair care.

**Topic Keyword View:** In order to reduce the issues with word coverage, this framework first extracts a list of keywords from the selected media topic. Next, the

analyst can explore the uniqueness of these keywords through the Topic Keyword View, Figure 4.2. In this view, a small multiples bar graph is used to show the 50 most frequent keywords. The height of the left bar represents the frequency of the word with respect to documents that are classified into the selected topic. The height of the right bar represents the frequency of the word in all other documents in the dataset. This view has two options for ordering these keywords, either by the word frequency in this topic or by the significance of this word with respect to all other topics in the dataset. The significance measure of a keyword  $w$  to a topic  $d$  is defined by:

$$\text{significance}(w, d) = \frac{f(w, d) - \sum_{t \in T, t \neq d} f(w, t)}{\sum_{t \in T} f(w, t)}, \quad (4.1)$$

where  $T$  denotes the set of all topics extracted from the media collection,  $f(w, t)$  is the frequency of word  $w$  in topic  $t$ , and  $d$  is the topic under analysis. The range of this metric is  $[-1, 1]$  where values closer to 1 means the word is more significant in the chosen topic. This measure is also perceptually visible based on the height of the two bars. If the left bar is taller than the right bar, the significance value is positive; if the left bar is shorter than the right bar, the significance value is negative. Analysts can select keywords by clicking on the bar graph and they will be highlighted by a rectangular box. In Figure 4.2, the five most frequent words in the topic, *Economy*, have been selected and are shown at the first line by the frequency order (Figure 4.2(a)). When the view is reordered based on the significance metric, three of the five most frequent keywords are also listed as the top five most significant keywords while the other two fall into the last ten words of the list (Figure 4.2(b)). Using this view, a keyword selection reference is provided so that the analysts can choose representative words for the media topic, thereby injecting domain knowledge into the semantic annotation pipeline while reducing the word set chosen for semantic

matching.

**Similarity Measure:** Once keywords are chosen, semantic matching to identify relevant links between the datasets is performed. A semantic similarity score is calculated between the selected keywords and the documents in the secondary dataset which is then filtered by this score and the documents that have a high semantic similarity score are returned for evaluation.

A knowledge-based word semantic similarity metric, Wu and Palmer [95], was used to first calculate the word-word similarity between selected media topic keywords and all words in the secondary dataset. This metric measures the depth of two given senses in the WordNet taxonomy [96], along with the depth of the least common subsumer (LCS). The sense-to-sense similarity is calculated as follow:

$$SenseSim = \frac{2 \times depth(LCS)}{depth(sense1) + depth(sense2)}$$

This is a sense-to-sense similarity measure, but it can be used as a word-to-word similarity measure by selecting the highest similarity score among all the similarities between the senses of these two words. Thus, word similarity can be defined as  $WordSim = Max(SenseSim)$ , which is a score between 0 and 1.

Given a keyword set representing the media topic and the word-to-word similarity measure, a semantic similarity metric can be developed to measure the relatedness of a document to this media topic. The media topic can be described as a set of keywords  $K = \{k_1, k_2, \dots, k_m\}$  where  $k_i$  is one of the  $m$  keywords. Similarly, the document in the secondary dataset can also be represented by a set of words  $E = \{w_1, w_2, \dots, w_n\}$  where  $w_i$  is the word occurring in the document from the secondary dataset. Note that both the media keywords in  $K$  and the words in the secondary dataset  $E$  are preprocessed to remove stop words and are lemmatized using CoreNLP [97] for con-

sistency. Using the above notations, our similarity score between a topic in dataset one and a document in dataset two is calculated as follows:

$$EventSim(E, K) = \sum_{\substack{w, \exists k_i \in K, \\ WordSim(k_i, w) > \theta}} \delta_w \text{tfidf}(w, E), \quad (4.2)$$

where  $\theta$  is a threshold to filter for semantically similar words (by inspection,  $\theta = 0.8$  was a reasonable choice and is used as the threshold values for all examples), the tf-idf is used to weight the word’s importance, and  $0 \leq \delta \leq 1$  is a weight for each semantically matched word. The value of  $\delta$  is initially set to be 1 for all words, and  $\delta$  can be changed during the visual to parametric interaction methods that are described in Section 4.2.2. The augmented frequency was used for  $\text{tf}(w, E)$  to prevent a bias towards longer documents, where

$$\begin{aligned} \text{tfidf}(w, E) &= \left( 0.5 + \frac{0.5 \times f_{w,E}}{\max\{f_{w',E} : w' \in E\}} \right) \times \text{idf}(w, D) \\ \text{idf}(w, D) &= \log \frac{N}{|\{E \in D : w \in E\}|}, \end{aligned}$$

and  $w$  is one word in a document of the secondary dataset  $E$ , and the whole secondary data collection is  $D$ , the size of which is  $N$ . The inverse document frequency,  $\text{idf}(w, D)$ , is logarithmically scaled. This approach returns a list of documents ordered by their similarity scores together with the words that are semantically similar to at least one of the media topic keywords. The semantically matched documents from the secondary dataset are shown in the Event List view (Figure 4.5) since in our analysis each document in the secondary dataset describes one event.

#### 4.2.2 Semantic Similarity Update

As previously mentioned, a direct calculation of semantic similarity from a knowledge base has issues with one keyword semantically belonging to multiple related concepts, for example, if you look at relationships for “food” in the knowledge base,

both “bread” and “education” appear; however, these represent two very different concepts which may not be the intention of the analyst. Thus, a visual to parametric interface is developed, building on the conceptual work of Leman et al. [98], in which the analyst can cluster the concepts returned from the knowledge base to better refine the semantic similarity matching. In addition to clustering concept words, the analysts can mark entries returned from the secondary dataset as relevant or irrelevant which will update the word similarity weight ( $\delta$  in Equation 4.2) thereby modifying the semantic scores and reorganizing the event list.

#### 4.2.3 *Semantic Interaction for Concept Word Clustering*

Again, even though a chosen keyword may semantically match a word in the secondary dataset, this matching may not align based on the contextual concept in which an analyst is working. For example, the word “food” has the following three noun senses:

- food#n#1: any substance that can be metabolized by an animal to give energy and build tissue;
- food#n#2: any solid substance (as opposed to liquid) that is used as a source of nourishment, and;
- food#n#3: anything that provides mental stimulus.

If an analyst wants to relate concepts of food and agriculture, sense 1 and 2 are likely related to the semantic search; however, sense 3 is unrelated.

#### **Cluster Force Layout:**

Based upon the interface design of IN-SPIRE [99] that uses word clusters to represent document themes, a cluster force layout (Figure 4.3) has been developed to

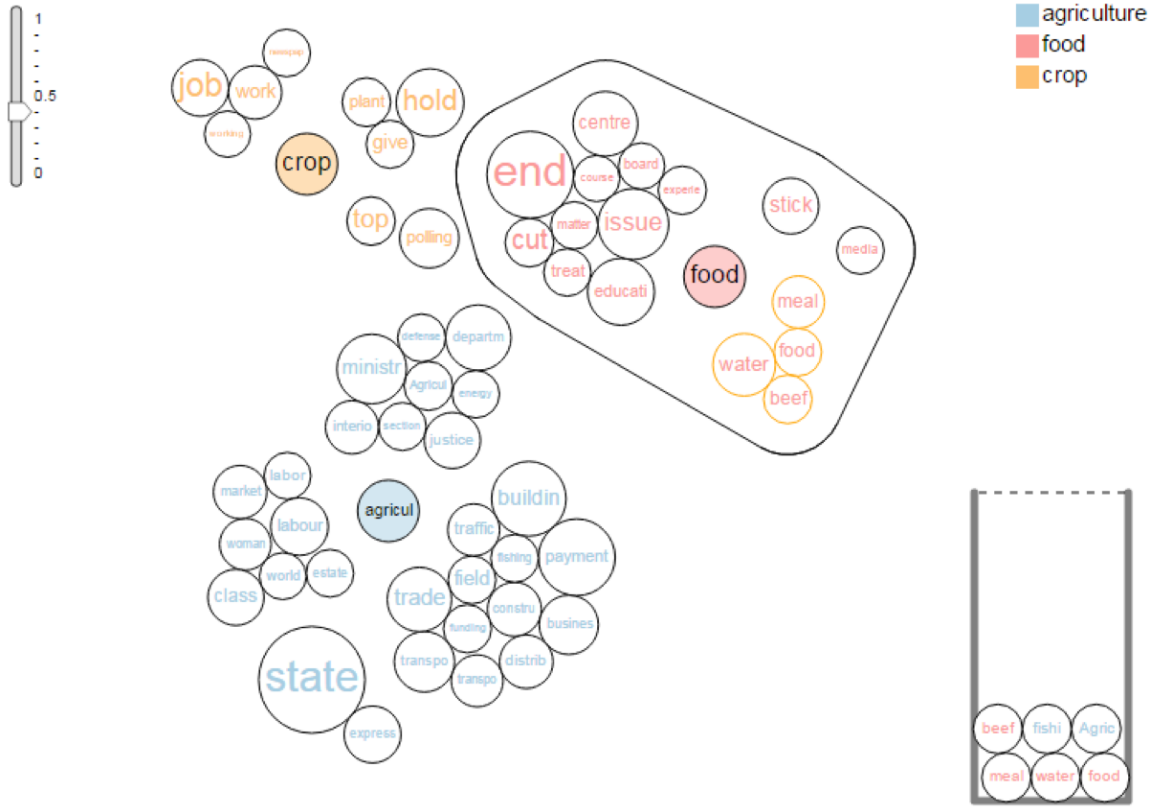


Figure 4.3: The Cluster Force Layout.

group words in the semantic dictionary based on their word-to-word similarities. The features include methods that enable the analyst to steer and update this clustering in order to develop an appropriate concept map for semantic annotation. To separate words by their meaning, complete-link agglomerative hierarchical clustering [100] is used, in which the similarity between two clusters is decided by the smallest similarity of all word pairs between the two clusters. To develop a concept map, an analyst can interact with the cluster force layout (Figure 4.1 bottom middle). In the layout, each node represents a word. The nodes with a solid background represent the selected keywords from the media data, and the nodes without a background color represent the words extracted from the secondary dataset that are semantically related to the selected keywords. This view uses a categorical color scheme to separate different



keyword bubble groups. Words in the secondary dataset are attracted to their corresponding keyword in the media data, and words belonging to the same cluster are further attracted together. Collision detection is applied to ensure that the nodes do not overlap each other, and nodes in different clusters are separated by a larger margin. As a result, each keyword and its semantically related words naturally form a bubble group, along with internal bubble clusters formed by the clusters of the semantically related words. We call the former the keyword bubble group and the latter the bubble cluster. The nodes in each keyword bubble group are colored to match the keyword legend on the top right corner. The size of each node is proportional to the frequency at which the word appears in the event records. In the case where a word becomes too small to see, the analyst can mouse over the nodes to show the words in a tooltip. Sometimes the analyst may select many keywords and the keywords may contain many semantically related words, then there will not have been enough space to display all the words in the view. In practice, a  $900 \times 450$  space for this view can support 4 keyword bubble groups with around 100 semantically similar word bubbles. To enable the capability of analyzing more words, zooming and panning on the cluster view is also allowed. The analyst can also freely drag the keyword nodes and the cluster nodes to adjust their relative position. As the clusters move close to each other, an attractive/repulsive force will be activated based on the similarity between the two clusters. This similarity score is calculated by taking the average of all word pair similarities. If the similarity between the two clusters is greater than 0.5, the clusters will attract each other, otherwise, they will repel each other.

**Implementation of Cluster Force Layout:** The force directed layout from the d3 library [101] was used. The clusterings were formed by adding a gravitational force to each of the selected keywords such that the keywords would only attract the words

that belong to their clusterings. Then, the clusters were formed by adding a force between words that belong to the same clusters. A collision detection force was also added to prevent the nodes from overlapping each other, and the collision detection force will separate nodes in different clusters by a larger space than nodes in the same clusters. To stabilize the force layout and prevent jittering during interaction, a repulsion force was also added between each node to neutralize the other attracting forces when the layout has already formed its shape to prevent the nodes from constantly colliding. However, when a cluster moves close to another cluster that belongs to a different clustering, the repulsive force would turn into an attractive force toward the nodes between the two clusters. Also, whenever the user drags any clusters on the layout, the strength of the attractive forces for the cluster will be slightly increased to ensure the cluster shape is preserved.

**Semantic Interaction:** In addition to visualizing word clusters, the cluster layout allows analysts to select sub-clusters and filter out words through semantic interactions [102–104]. The underlying concept is that by allowing users to directly manipulate data in the visualization space, updates to the positions of data elements on the screen can be tied back to weights in the analytic modules on the backend, which can then be translated to the model updates. Our cluster force layout supports semantic interactions for creating concept clusters. Here, a user can change the number of clusters by dragging the slider on the top left corner to set the similarity threshold in the hierarchical clustering and change the similarities between words through drag and drop interactions on the bubbles. Let  $k$  denote the number of clusters shown in one keyword bubble group, and its word set can then be represented by clusters  $C_i = \{w_{ij}\}$ , where  $i = 1, 2, \dots, k$ ,  $j = 1, 2, \dots, n_i$ , such that cluster  $C_i$  contains  $n_i$  words. An analyst can drag a word,  $w$ , from its current cluster  $C_i$  to another cluster

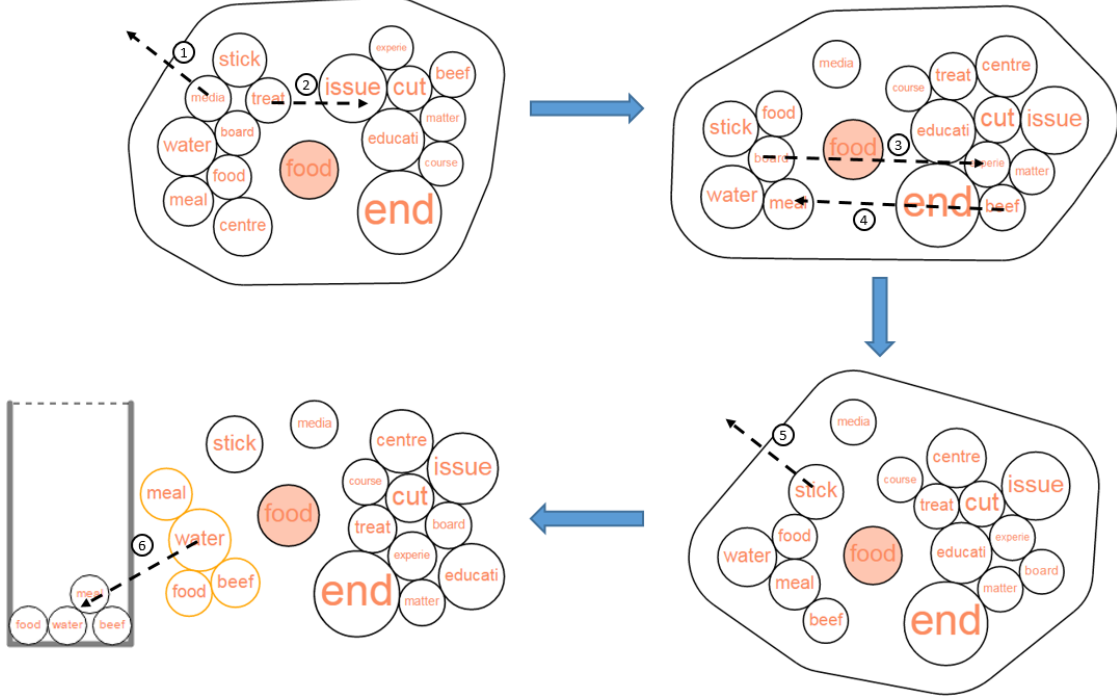


Figure 4.4: This Figure Shows an Interaction Process of Steering the Cluster Force Layout to Update the Word Concept Clusters. Each Step Is Marked Numerically on the Figure. User Can Drag a Word Away from All Clusters or Towards Another Cluster. User Can Drag a Group of Words to Select Them.

$C_{i'}$  such that the similarities between  $w$  to all other words in  $C_i$  decrease and the similarities between  $w$  to all words in  $C_{i'}$  increase while the similarities between  $w$  to the words not in  $C_i$  nor  $C_{i'}$  do not change. The new similarities of word  $w$  to other words updates as follows:

$$sim'(w, w') = \begin{cases} sim(w, w') + (1 - sim(w, w')) \times 0.1, & w' \in C_{i'} \\ sim(w, w') \times 0.9, & w' \in C_i \\ sim(w, w'), & otherwise \end{cases}$$

Here  $sim(w, w')$  represents the similarity between the word  $w$  and  $w'$  before the interaction, and  $sim'(w, w')$  represents the new similarity after the interaction. If

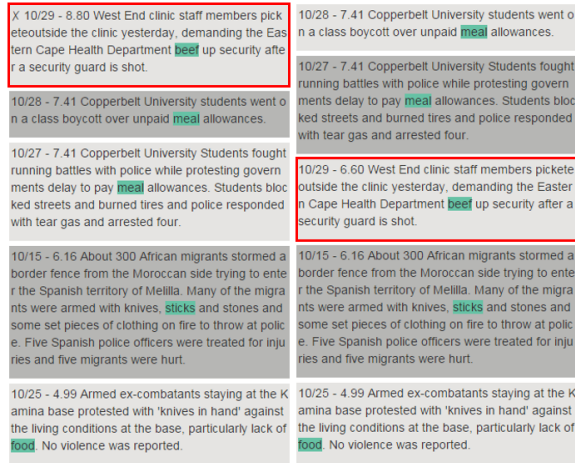


Figure 4.5: Interaction on the Event List View to Update Word Weight and the Event Semantic Similarity.

the analyst finds that there are no clusters that the word  $w$  can join, s/he can try to reduce the similarity between this word and all the other words in the keyword bubble group. To do this, s/he can click on any nodes in this keyword bubble group to activate the convex boundary and then drag the word  $w$  outside the boundary. Doing so will reduce the similarity between the word and other words in this keyword's group by 50%. The change of these similarities will trigger an update of the hierarchical clustering and the cluster force layout. To hide the boundary, the analyst can click any of the nodes in the bubble.

Once an analyst is satisfied with a concept cluster, s/he can choose the words in the cluster to be used for semantic similarity matching for the event records. The analyst can select the cluster by holding the mouse on any of the nodes in the cluster, and the selected cluster will be highlighted by an orange border. The analyst can then drag this cluster into the container in the bottom right corner to select the words, and those selected words will remain in the container and be used for semantic similarity matching. Alternatively, the analyst can also drag any individual words to

the container. Selected words can be removed by dragging them out of the container. An example of these interactions is illustrated in Figure 4.4 which shows how we can use the cluster force layout to eventually select a subset of words related to “food”. In Figure 4.4, the analyst is creating a concept map for the keyword “food”. The bubbles contain semantically related words as captured using the WordNet similarity. First, the analyst inspects the different clusters. In step 1, the analyst wants to refine the cluster containing “meal”, “food”, and “water”. The word “treat” is moved into the other cluster within the keyword bubble group and the clustering updates. Due to the semantic similarity score between “treat” and “centre”, “centre” is also moved with “treat”. In step 2, the analyst wants to remove “media” entirely from the analysis and drags it outside the convex boundary of the “food” clustering. Then, in step 3, “board” is moved away from “food” but positioned next to “cut” as the analyst feels those may be conceptually similar. After having a cluster with words “food, water, meal”, the analyst notices that the word “beef” might also relate to the concept of food, so in step 4, “beef” is dragged to the cluster of food. This turns into a state where the word “stick” is also clustered together with food and the analyst drags it away as shown in step 5. Finally, in step 6, the analyst chooses to use the words in the highlighted cluster for the semantic matching.

#### 4.2.4 *Word Weights Update*

In addition to semantically interacting with the cluster force layout to refine the concept words, users can also interact with the Event List showing the text details from elements in the secondary dataset (Figure 4.1 bottom right). Each record in the Event List view contains the date, similarity score, and the text for an event. The selected semantically similar words are highlighted using the same color as the related media keywords. For example, the Event List in Figure 4.1 shows events queried by

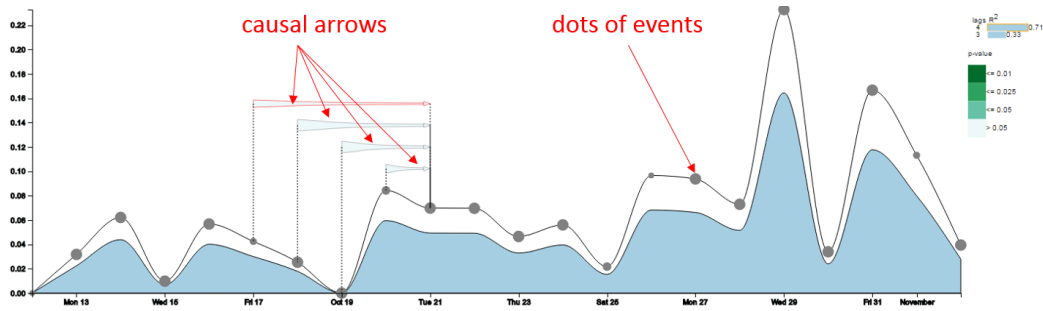


Figure 4.6: Initial Causality Results Without Filtering Events. The Causal Arrows Are Colored Based on Their Significance, and the Current Result Shows No Significant Causality Between the Current Media Stream and Events, and the Best Fit Model Is Displayed with  $R^2 \approx 0.6$  and  $0.1 < p\text{-value} \leq 0.5$  at a Lag of 4. The Height of the Area Curve Represents the Amount of Variance Explained by the Model.

topic keywords “food” and “agriculture”, which match to qualitative colors “pink” and “blue”. The semantically similar words (e.g. “meal” is related to “food” and “fishing” is related to “agriculture”) are highlighted by the corresponding color in the Event List. When browsing related events, the analyst can mark an event as relevant or irrelevant by clicking directly on the text. When marked as irrelevant, the word weight  $\delta$  in Equation 4.2 will decrease by 0.25 (until reaching 0), and  $\delta$  will increase by 0.25 (until reaching 1.0) if marked as relevant. Through this interaction, the scoring measure of the events will update while the word similarity cluster does not change. For example, in Figure 4.5, we filter a list of events based on words similar to food, and we notice that the word “beef” in the first event does not mean the meat for eating but means “to strengthen”, and we mark this event as irrelevant. The weight of the word “beef” then decreased and the rank of this event drops, as shown in the right side list.

### 4.3 Causality Detection

While the semantic similarity analysis and interactions enable the analyst to filter and link events between two datasets, these methods do not provide any indication of whether or not the events identified in the secondary dataset seem to be driving the media topics under analysis. As such, this framework leverages statistical causality models to provide a quantitative indicator of significance under the hypothesis that the current event series is driving the media discourse. The input to the causality model is formulated as two time series. Time series  $Y(T) = \{y_1, y_2, \dots, y_t, \dots, y_T\}$  represents the volume per time step of documents in our media data classified into the topic of interest. Time series  $X(T) = \{x_1, x_2, \dots, x_t, \dots, x_T\}$  represents the volume per time step of related events identified during the semantic matching procedure. Then, causality between these two time series can be tested for, where  $Y$  is the effect and  $X$  is the cause. Here, it is important to note that there may be other relevant factors in a larger universe to  $X$  and  $Y$  which cannot be modeled practically, so spurious causalities may also be identified; however, these measures are still able to provide insight and help in the hypothesis generation, testing, and exploration process. While no statistical technique can provide a definitive test for causality, a causality test is able to provide explanations of effects as the results of potential causes and suggest whether a change in the media stream might be correlated to some local events [105]. Another issue in this test is that the causal effect can be bidirectional, which means  $X$  can cause  $Y$  and  $Y$  may also cause  $X$ . In such a situation, a feedback mechanism should appear. In this application, we focus only on one directional causality, exploring the question of  $X$  causes  $Y$ . The following assumptions on our dataset are also made:

1. The cause shall appear before of the effect;

2. The information in a larger universe not coded in  $U = \{X, Y\}$  will be irrelevant, and;
3. Both  $X$  and  $Y$  are stationary series, which means their means neither change over time nor follow any trends. This should be reasonable for natural events, otherwise they shall first be transformed to stationary processes.

The **Granger causality test** [32] is applied. In a simple causal model (no instantaneous causality and no feedback mechanism), causality is tested by fitting the following two linear regression models and testing if the prediction variance is statistically significantly improved in the second model.

$$y_t = a_0 + \sum_{i=1}^m a_i y_{t-i} + \varepsilon_t$$

$$y_t = a_0 + \sum_{i=1}^m a_i y_{t-i} + \sum_{i=1}^m a_i x_{t-i} + \varepsilon_t$$

Here,  $\varepsilon_t$  is an uncorrelated noise series, i.e.,  $E[\varepsilon_t \varepsilon_s] = 0, s \neq t$ , and  $m$  can be any integer in  $[1, T]$ .

Let  $\sigma^2(A|B)$  be the variance of  $\varepsilon_t(A|B)$  which is the error series in the prediction model that series  $A$  is the response and series  $B$  is the predictor. In the test, given a value of  $m$ , it can be said that  $X$  Granger-causes  $Y$  at lag  $m$  is statistically significant if

$$\sigma^2(Y|Y - Y(t - m), X - X(t - m)) < \sigma^2(Y|Y - Y(t - m)). \quad (4.3)$$

The F-test is used here to test the significance of the increment of the explanatory power of adding  $X$  by comparing the overall fit of the model using only  $Y$  and then by using both  $Y$  and  $X$ . A corresponding *p-value* is also used to show the significance level, where the null hypothesis is that the variance has not decreased by adding  $X$ . An  $R^2$  value is used to indicate the performance of the second model by showing how much variance can be fit using both  $X$  and  $Y$ . The causality test treats the



two timeseries data as two arrays of data, the length of time gaps between each data point depends on the granularity of the timeseries. Currently in our system, the granularity of our data is in days, but the causality test will also work for hourly or monthly data. Since the question of “true causality” requires field testing and controlled experiments, the applied statistical method should only be considered as “predictive” causality which tests whether one time series is useful in forecasting another [106]. However, this is useful as an indicator for trends and drivers and can help an analyst in exploring hypotheses. This serves as a basis for choosing which points in the Timeline may need further annotation.

#### 4.4 Annotation

Annotations have been used in visualizations to highlight interesting data points, provide context, and display detected events [88, 107–109]. Our system allows analysts to annotate media articles and related events, and provides causality modeling indicators in the Timeline for correlation discovery and externalization.

##### 4.4.1 *Timeline with Events and Causalities*

One of the main views in this system is the Timeline, Figure 4.6, which starts from a single line denoting the volume of the media stream to be augmented gradually by adding relevant events, causalities, and descriptive text annotations. The solid black line indicates the trend of the percentage volume of the selected media topic. The dots in gray on this line represent related events that happened on a given day, and the size of a gray dot represents the number of events that happened on that day. The size and the amount of the dots will update after interactions with the Topic Keyword view, the Cluster Force Layout, and the Event List view.

When the analyst is satisfied that the relevant events have been semantically linked, they can click on the “Causality Test” button (see Figure 4.1) to run the causality test on the retrieved events and the media topic. This causality test returns the statistics for all models with possible lag smaller than 10. For each model, the *p-value* and  $R^2$  are displayed for evaluation. These causality models with different lags are indicated at the top right corner of the Timeline by a legend consisting of several bars, one for each model. To the left of each bar, the lag of the model is shown, and the length of the bar indicates the  $R^2$  value. When one model is selected, a stream area, referred to as the explanatory area, is shown below the topic volume line to represent the  $R^2$  which denotes how much variance is explained by the model. For example, when a model’s  $R^2 = 0.6$  the explanatory area will cover 60% of the area under the media stream line, Figure 4.6. The analyst can move the mouse along the timeline to browse the model for each date, and the identified lags and events will be shown on top of the timeline illustrated with arrows (causal arrows in Figure 4.6) connecting the possible driving events to the effect date of media articles. The color gradient, in a sequential color scheme, shows the respective significance levels referring to the *p-value* color legend. For example, in the Figure 4.6 the color of the causal arrows are light green which means the causality is only significant at a level lower than 90%.

For a model with lag  $m$ , the media stream is fitted with a linear regression model  $y_t = a_0 + \sum_{i=1}^m a_i y_{t-i} + \sum_{j=1}^m b_j x_{t-j}$  and only the past  $m$  time steps have been used as predictors. The start point of each arrow matches the time of the event and the end point (the point with an arrow) matches the time  $t$  where the volume of news articles are being predicted. The width of an arrow’s starting point corresponds to the amount of events happening that day. For the purpose of perception and aesthetics, the arrows are ordered by their length either bottom up or top down, based on the

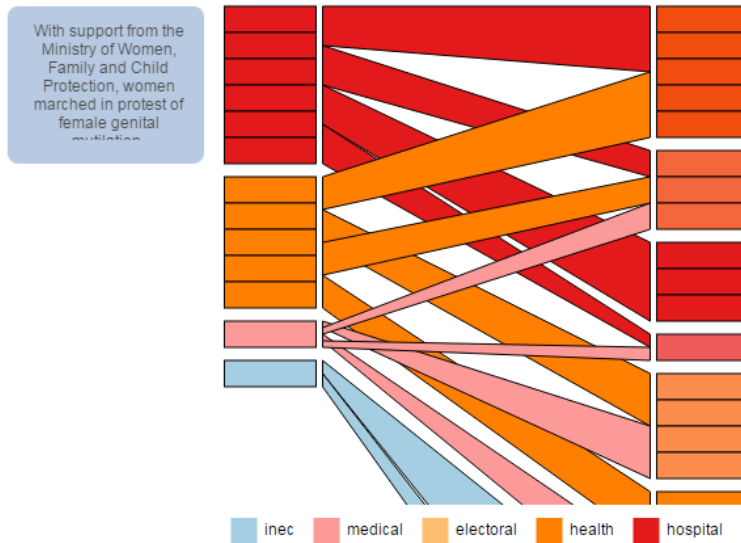


Figure 4.7: Example of the Bipartite View. This View Shows the Events (Left Side) and the Media Articles (Right Side) Indicated by the Causality Lag Under Analysis and Connected by Semantically Matched Keywords.

position of the gray dots which represent the potential causal events. Our timeline view currently supports only the results from one causality test on one pair of series, and future work will explore methods to overcome this limitation.

#### 4.4.2 Annotating with Text Information

When the causality test result is positive, it shows the possibility that event series is the cause of the media series, however, to make such hypothesis it is necessary to look into the content of the media articles and the detail information of those events. A bipartite view is used to allow the analysts to navigate through the events and media articles, and then the analysts can pick interesting events and media articles to be annotated on the timeline.

### 4.4.3 Bipartite View

The Bipartite View (Figure 4.7) displays the connections between media articles and event records linked by the causal arrow. The bipartite view updates while the analysts anchor on a date to investigate (by moving the mouse on the timeline while holding the left key). Initially the bipartite view displays the events and media articles on the same day. Once the analyst selects a causal arrow by clicking it, the bipartite view displays the events and media articles linked by the arrow. The right side of the Bipartite View lists all media articles and the left side lists relevant events. Both media articles and events are colored according to the semantically matched keywords. For media articles, the semantically matched keywords are the selected media keywords that the articles contain. For events, the matched keywords are the selected media keywords that have some of their semantically related words contained in the event notes. When the article or event record has one matched keyword, it will be represented as a rectangle in the color corresponding to the keyword's legend. If more than one keyword has been matched, this rectangle will use a blended color from all the matched keywords' colors. Both media articles and events are grouped according to their color, i.e. the semantically matched keywords. The edges connect the groups by their keyword co-occurrence. The edges are colored by the color of matched keywords that both of the connected groups contain. For example, Figure 4.7 shows the bipartite view of topic "Election" and keywords "inec" (Independent National Electoral Commission), "medical", "electoral", "health", and "hospital". Mousing over any rectangles on the bipartite view will bring out the tooltip which shows the beginning of the event note/article. For displaying a long list of matches, the Bipartite View is scrollable so that the user can view the details when they do not fit in the area.

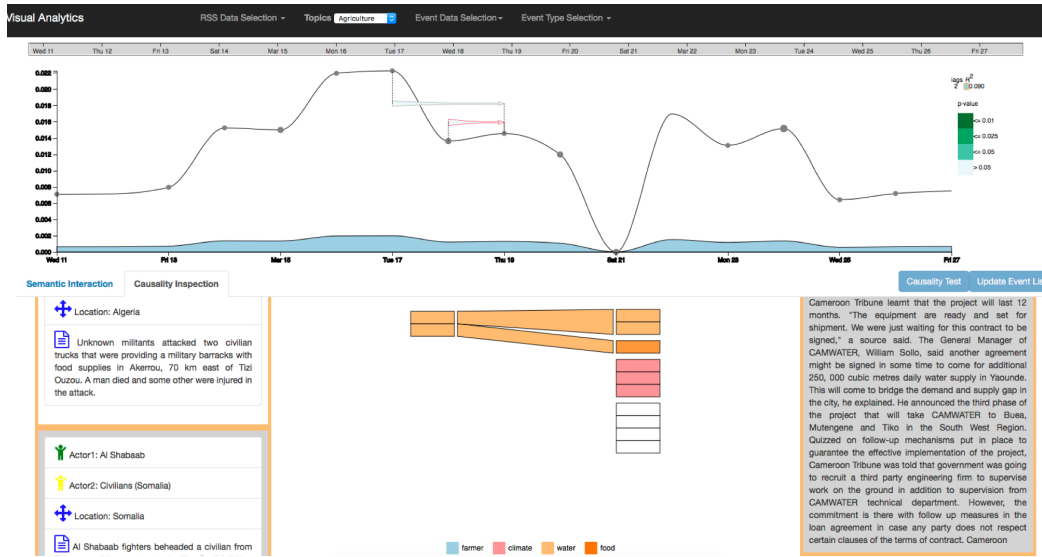
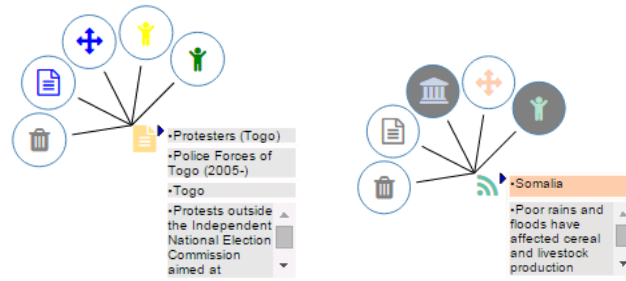


Figure 4.8: Investigating the Plausibility of Climate-Induced Civilian Abuse during the 2014 GHoA Drought. The Link between Violence Events and Social Unrest RSS Have Been Explored Although Casualty Test Shows Non-Significant Result.

Through the annotation method described above, the analyst can build a story talking about correlations between the media topic and the possible causing events with a timeline depicted as Figure 4.1.

#### 4.4.4 Detail List

The detailed information that the rectangles in the Bipartite View contain can be explored in the linked detail view on the sides of the Bipartite view. The detail of event records are listed on the left and the detail of media articles are listed on the right (Figure 4.8, bottom). Each event record has up to two actors, a location, and a note describing the event. The border color of each entry in the lists matches the color of its matched keywords. Different to the color design of the Bipartite View, if there are more than one matched keyword for the entry, there will be multiple borders,





(a) Event Annotation      (b) Media Article Annotation

Figure 4.9: The Annotation Glyph for Event and Media Articles. The Analyst can Annotate Entities by Clicking on the Expanded Nodes.

each colored by one keyword’s color. The order of entries on both lists reflects the order of the rectangles on the bipartite view.

#### 4.4.5 Entity Annotations

While exploring the detailed information through the bipartite view, the analysts can externalize their findings to the timeline by double clicking on the media and event rectangles. The selected event will be annotated as a record icon  using its keyword’s color. The selected media text will be annotated as a feed icon  using the color of its keyword. To expand the detailed information of the annotated events and media articles, analysts can click on an icon, and the actors, locations, and text information will be available as expanded nodes (Figure 4.9). From these nodes, the analyst can choose which event attribute to show on the Timeline. The annotation can help analysts to immediately interact with data so that one can flag events that can constitute changes in the underlying equilibrium of these processes.

## 4.5 Usage Scenario

In this section, we demonstrate this system through an analysis using a climate change media collection and a social unrest media collection respectively. These datasets will be semantically annotated by the Armed Conflict Event Location Dataset (ACLED) [110] and causal drivers explored. For the use cases, a paired analysis protocol [111] was used in which system features were explained and demonstrated to our partners in political science. The analysts discussed their developing hypotheses and instructed the framework developers driving the system.

### 4.5.1 Datasets

**ACLED:** The ACLED dataset (1997 to present) contains information on the dates and locations of all reported political violence events in over 50 developing countries with a focus on Africa. Each event record contains information on the date, location, event type and actors involved with approximately 6500 events from August to December 2014.

**Climate Change Media:** The climate change media dataset is composed of RSS feeds from 122 English language news outlets and filtered for relevance by matching against a set of 222 keywords. From August 2014 to December 2014, this collection contains 1245 relevant articles with 9070 sentences which are further coded into one (or none) of 25 framing categories. All articles have been analyzed through entity recognition to extract people, location, and organizations. A more specific description can be found in [90].

**Social Unrest Media:** The social unrest media dataset is composed of RSS feeds from 128 English language news outlets collected in March 2015. RSS feeds were scanned hourly and the content of each news article was filtered by a set of 378 social

unrest keywords. The LDA topic modeling algorithm [112] was run on these articles and 50 topics were extracted. The following 7 topics were selected based on their relatedness to the ACLED dataset: Election, Economy, Education, Conflicts, Agriculture, Justice, and Energy. All articles have been processed using entity recognition for annotation.

#### 4.5.2 *Climate-induced Unrest During Drought*

The drought in Africa has attracted the attention of researchers who want to analyze potential societal impacts that the draught may have [113]. Specifically, the draught has caused widespread agricultural failures and led to famines and political instability. In this use case, the analyst wanted to explore if the GHoA (Great Horn of Africa) drought in 2014 coincided with instances of political violence. Causal relationships are extremely difficult to test when using observational data, and political scientists debate about the relationships between climate change and political violence [114]. Furthermore, the research that examines the role of drought in instances of violent conflict lacks consensus [115], and it remains unclear whether the onset of drought and subsequent observations of violence could be indicative of broader phenomena, such as a governance failure or institutional failure. As such, this use case focuses on the question: Is the 2014 drought in the GHoA linked to reports of social unrest and political violence? To probe this question, the researcher first selected “social unrest RSS”, defined the temporal domain to encompass the month of March 2014, and picked agriculture as the topic to explore. Next, events of “violence against civilians” (which ACLED defines as any armed/violent group that attacks civilians”) is selected to explore potential drivers between droughts and violence.

To begin, the analyst selected the following words, based on the significance ordering, as most plausibly being related to the March - June 2014 GHoA drought:



“water”, “food”, “farmer”, and “climate”. These selections are then used to update the relevant events and explore the semantic linkages in the clustering bubble interface displayed below the event timeline. The clustering results reveal that the word “climate”, in this semantic mapping, shares words/concepts not related to the concept of agriculture. Instead, words such as “way, order, demand, tension, and control” are found, even after adjusting the similarity threshold to .75. Thus, “climate” is removed and the event list is updated to reflect the change (see Figure 4.10). After removing “climate”, the analyst remains largely satisfied with the clustering for the remaining terms and now creates clustering within the keyword selection container. The causality test is performed and returns the following insignificant model result: lag=2, R2=0.090 (see Figure 4.8). The insignificant result is not surprising to the analyst since many other factors are also expected to drive the events. However, he also requested to further explore the details of the events and the media posts to identify if there were other keywords or factors he may not have considered. Using the Bipartite View, the analyst briefly evaluated the links between the keywords and the recorded episodes of political violence perpetuated against civilians. His search revealed shared associations between terms related to drought and a recorded event of Al Shabaab beheading a civilian for unstated reasons, (see Figure 4.8). Based on the exploration, the analyst concludes that the linkage between resource shortages and civilian abuse may be less plausible.

### *4.5.3 Food Insecurity and Climate Change Media*

The analyst was interested in exploring potential drivers of climate change media discussions with respect to ongoing conflict events in Africa. He hypothesized that external drivers, such as riots and protests, may be driving the types of framing being used to discuss climate change. First, the analyst loads the Climate Change

media collection and the ACLED dataset. The analyst decides to focus on the food insecurity frame “ProbThreatFood”, from October 12th to November 3rd, 2014. Next the analyst chooses to explore “Riots/Protests” from ACLED to annotate the media frame. The analyst first selects keywords “food”, “crop”, and “agriculture”. The analyst adjusts the threshold in the Cluster View and discusses the resulting clusters.

First the analyst chooses several topic keywords related to food insecurity in the climate change media dataset, selecting “crop”, “food”, and “agriculture”. Events are then automatically selected through the semantic keyword processing, and the analyst runs an initial causality test which shows no significant causal correlations. The result of this initial model is shown in Figure 4.6. The problem is that many events that are marked as semantically similar to these keywords do not match the analyst’s meaning of “crop”, “food”, and “agriculture”. Thus, the analyst begins using the Cluster View to group the words into conceptual groups. As such, the analyst explores semantic keywords related to the selection of “crop”. However, the analyst finds that events in the secondary dataset matched as semantically similar to “crop” are not embedding the meaning of “cultivated plants”. Instead, the related words, particularly “plant”, are referring to industrial plants in the event data, but not cultivated plants. This illustrates the difficulties that a fully automatic approach might face. While it may be reasonable to assume that “plant” as a keyword would represent agricultural concepts, in the ACLED dataset this is not the case. Thus, the analyst removes “crop” from the keyword list so that events semantically linked to this concept of crop are no longer selected. Again, since different domains use slightly different linguistic descriptors, a visual analytics solution can enable analysts to inject domain knowledge and reasoning into the analysis pipeline.

Next, the analyst explores the clusters of semantic keywords linked to “food”, as shown in Figure 4.4, and conceptualizes the meaning of several clusters. Seeing

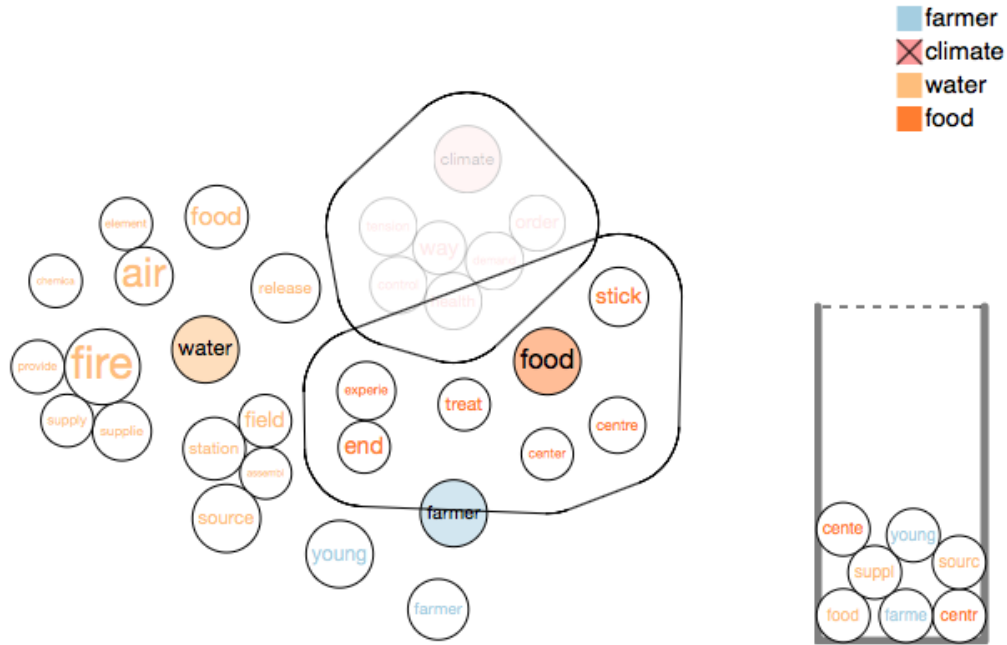


Figure 4.10: Semantic Word Clustering and Filtering for the Study of Climate-Induced Civilian Abuse During the 2014 GHoA Drought.

words in the blue cluster, such as “education”, “course”, and “issue”, being clustered together was identified as primarily being a concept of “food for thought”. The analyst finds that one cluster contains many “food” related words (e.g. meal, water, food). However, there are still some irrelevant words in this cluster (e.g. treat, centre, board and media). To better select related events to “food”, the analyst drags these irrelevant words either into the other cluster or away from both clusters, the result of which is shown in Figure 4.4 - Step 4, where one cluster now contains “food”, “water”, “meal” and “beef”. Now, the analyst chooses to only semantically link the word “food” to match his conceptual notion.

The analyst then chooses to filter only on certain conceptual cluster groups that he has now created in the cluster view. For example, in the cluster view illustrated

in Figure 4.4(step 6), the cluster with words (food, water, meal, beef) are selected by dragging into the word selection container and used to semantically link the media dataset to events. However, after exploring the event notes, the analyst realizes that “beef” is only used in the context of “beef up”, so the analyst removes the word “beef” from the semantic filtering as this meaning does not relate to the current conceptualization of food. The analyst then further filters the events using his agriculture concept group. After filtering events, the best fit causality model at this point is still not significant ( $R^2 \approx 0.6$  and  $p\text{-value} > 0.1$ ), Figure 4.6. From the event list view, the analyst sees that there are many words in the event data that are semantically linked by the media keyword “agriculture” but not related to the meaning in the context of food problems. The analyst again refines the Cluster View to select relevant words for “agriculture” and selects “agriculture” and “fishing” (as Africa has large exports of fish) which is shown in the word selection container in Figure 4.1. The causality modeling returns three models with lag equals to 2, 3, and 4 respectively. The best fit model (Figure 4.1) has lag = 4,  $R^2 \approx 0.8$  and  $p\text{-value} \leq 0.05$ .

Switching to the Causality Inspection tab for more details of the relevant media articles and ACLED events, the analyst explores the connections suggested by the causality model with lag = 4 using the Bipartite view to further annotate potential causal relationships. Here, the analyst explores filtered events, annotating nodes with text and exploring lags in the dataset. The final resulting annotated Timeline that was developed is shown in Figure 4.1. In looking at the peak discussion of food insecurity, we are seeing articles mentioning Food shortages and rising prices have the potential to worsen political, ethnic, class and religious tensions, and “Ordinary people could go hungry if their countries can not produce enough food”. Prior to these articles, we are seeing an increased number of riots and protest around unpaid meal allowances and discussions of lack of food. Here, discussions of riots and protests related to food

are marked on the Timeline and can provide insight into what external events may cause shifts in media framing, for example, if there are more protests related to food issues, does climate change media pick up on this and begin framing climate change as a food insecurity issue?

#### 4.5.4 *User Feedback*

The user feedback was collected from three groups of experts, a political scientist, the collaborators in communication, and our outside partners. The political scientist expert was interested in developing a system to enhance his data exploration process. This expert provided detailed feedback about tool usage and details in the use case are from an ongoing analysis. The second group consists of faculty and doctoral students in the Communications Department at Arizona State University. Their interest is in the impacts of framing in social media, which requires the ability to map different lexicons among topics. This group provided detailed feedback on keyword selection and analysis and used the system to perform social media analyses. The final group consists of experts whose feedback was informally solicited during demo sessions of the tool. Here, experts were walked through the tool usage and use cases, and discussions were held regarding how such a tool would be useful for their domain applications.

Through using the system themselves, the political science and communication experts liked the system overall, indicating that the major advantages were that this “Allows me to read in a way that I can’t do manually. This tool allows me to explore my beliefs about the data and events and record details to the point of theory construction.” Our experts liked the clustering force directed layout and consider it to be a more intuitive display. They felt it increased their understanding and willingness to engage with the program. The political science analyst also felt that the ability to easily hone the selection of the terms of interest by removing and adding

words to more relevant clusters is a striking feature, and the immediate updating of the clustering algorithm within the user-interface after winnowing the clustering to domain-relevant terms is very useful and appropriate. The analyst considers the embedded Granger causality testing to be simple to use, and the causality testing helped lead to formalization of actual hypotheses and provides some level of base knowledge about what concepts of interest are most relevant. The visualization of how much variance the model explained being represented as a filled area under the curve was noted as being highly intuitive.

Along with the above detailed feedback, the system has also been demonstrated to industrial partners. Feedback from these demonstrations indicated that users like the interface and the approach of semantically linking events to media topics. They think the visual representation of the clustered keywords are quite intuitive and the causality test is easy to understand. They also pointed out some limitations of the system. First, the system needs the text dataset to be preprocessed, including categorization (e.g. labeling by domain experts or topic modeling) and word similarity calculation. This limitation currently prevents this system from handling streaming text data. Second, the demonstrations only showed how to link between media posts and conflict events. However, these demonstrations were given to people in vastly different domains who indicated a need to analyze proprietary data sources which may require modifications to the visual design to support domain specific annotations.

Furthermore, many media sources of interest also contain video and images, as such, extracting relevant content becomes difficult. Also, the scalability of the system is a critical issue. In the datasets used by the collaborators for the case studies, scalability was not an issue as some data curating had already been performed and pre-processing of data could take place prior to interactive analysis. However, a key task of causality analysis is to build predictive models. During our demonstrations,

requests for real-time model building and updates using streaming data was discussed. Currently, the system is limited by the pre-processing requirements; however, the workflow proposed by the system is robust to support the causality modeling task but will require the addition of a streaming data processing step.

#### 4.6 Discussion

This case study presents a system for semantically annotating media topic discourse through linked datasets. To accomplish this, a cluster force layout that can facilitate the development of a concept map of keywords has been designed to be used for semantically filtering linked events. Relationships between these events and media trends can be analyzed using causality modeling, and model results are interactively displayed on the Timeline. Though the causality modeling cannot guarantee a true cause-effect relationship, results obtained from such models are able to help analysts in their knowledge discovery and hypothesis generation. Analysts may explore suggested connections between media articles and linked events, and articles and events that are linked by multiple concepts are further visualized in the Bipartite View. From the Bipartite View, analysts can annotate events of interest on the Timeline to help inform their given hypotheses.

While the usage examples focused on media and conflicts in Africa, the tools developed are applicable to a variety of domains and data. However, there are several limitations to this system. First, the semantic match is constrained to a keyword based approach, i.e., the analyst must choose an initial set of keywords from the document as a starting point. This can limit the matching as other words between the corpuses may serve as more appropriate semantic bridges. Second, although we have shown that the knowledge-based semantic similarity methods can be leveraged to connect two textual datasets, other information retrieval metrics could also be

explored and compared. Third, the system has limitations of its scalability and capability of generalizing to streaming data and more complexed data format. The scalability issue exists in both the cluster view and the Timeline as discussed. More advanced techniques in database and similarity calculations are needed to generalize this system for streaming data and broader datasets.



SPATIOTEMPORAL TRADE NETWORK ANALYSIS

Economic globalization connects regions of the world and creates complex interdependencies among countries. Through trade connectivity, the risks associated with climate-induced changes in agricultural production (e.g. floods and droughts) are exported from areas with localized disruption to distant parts of the globe [116]. For instance, the monsoons of Southeast Asia may seem to have little to do with food supplies in distant parts of the world. However, when one considers that countries as distant as St. Kitts in the Caribbean and Congo in Africa both rely on Thailand for over 95% of their imported rice, it becomes clear that a significant disruption in Thailand's agricultural production could have a dire impact on food security thousands of miles away. Without an appreciation of the effects of global networks, attempts to mitigate social and economic disruption may lead to investment, for instance, in countries expecting drought and not in countries that will face food shortages due to that drought. However, food and water vulnerability are typically forecasted for demarcated and isolated units – regions, countries, subnational units, cities – while ignoring the impact of connections to other units (e.g. [117]). Little has been done on how a trade network topology contributes to the delocalized vulnerability of food and water delivery systems, insights that are crucial to planning for sustainable and resilient access to food and water. Furthermore, it is important to understand how trade network topology and network-induced food risks contribute to distant political instability and social unrest [118–120]. While the network of global trade has been well-studied, those studies are largely limited to characterizing the magnitude of exports between countries [121–124], the quantities of virtual water transmitted

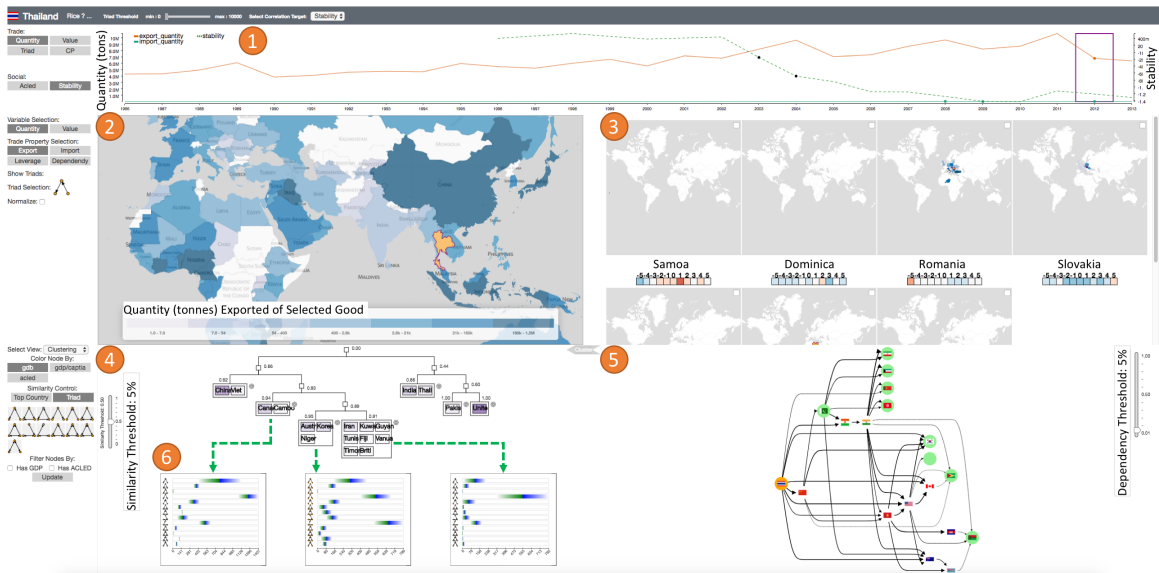


Figure 5.1: A Visual Analytics System for Exploring Global Trade Networks and Their Relationship to Regional Instability. The Anomaly Time Series (1) on Top Displays the Time Series and Anomalies of the Trade Attributes and Stability Measures for a Selected Country. The Choropleth View (2) Along with the Small Multiple Maps (3) Display the First Order Trade Relationships Centering on Selected Countries. The Colored Bars Below Each Small Multiple Map Show the Temporal Correlation of the Selected Trade Measure to the Stability Measure. The Clustering View (4) Displays the Hierarchical Clustering for the Countries, Based on Either the Triadic Similarity or Top Partner Similarity. The Groups Can Also Be Configured to Show the Average Triad Distributions (As (6)) and Other Measures. The Trade Diffusion Graph (5) Displays the Propagation Effect of Anomalies. Connections Between the Nodes Indicate the Import Dependency From The Target Node to the Source Node Is Larger Than a Threshold, Which Can Be Adjusted Using the Slider on the Right.

through those exports [125–127], or identifying the largest and most connected nodes in the network [128–131].

Currently, what is needed are tools and methodologies that enable users to explore complex aspects of local trade structures and their relationships to local vulnerabilities. Recently, the visual analytics community has begun developing tools for analyzing complex global events. For example, geo-social relationships to the global trade network have been visually explored [132], and the relationship between climate change and human conflict has been analyzed through linked geographical visualizations and statistical analytic views [90, 133]. Other analyses have called attention to how human migration might affect international trade [134] and how international trade may contribute to national air pollution [135]. Current methods have focused on geographic flow relationships [136, 137], movement of individuals and ships [138, 139], and spatiotemporal event detection [140, 141]. However, current methods have been limited to single temporal snapshots of flow and anomaly detection on space-time, ignoring networked effects. Understanding how trade relationships impact local vulnerabilities over time requires new methods that can provide an integrated analysis of local disruptions.

In the context of geographic network and flow analysis, common visualization methods quickly succumb to issues of overplotting. Geographic relationships need to be maintained to enable quickly filtering and exploring country relationships within a familiar spatial representation. However, as the number of attributes per country becomes large, and as the number of flows under analysis increase, methods for identifying data relationships and anomalies becomes critical. As such, this problem lends itself to a visual analytics approach where an analyst is needed to contextualize the geopolitical relationships of trade while being able to quickly cluster, filter and explore the data space. The contribution of this case study include a visual analytics system for combining multiple existing visualization techniques to help complement the user in processing large geographical data and their topological network relation-

ships. This case study introduces the concept of exploring triadic closures [14] over time, and by linking various triadic analysis features (a detailed explanation of triad analysis is provided in Section 5.1.3) with anomaly detection, the system enables advanced filtering to help analysts identify unexpected patterns over feature space, time and network topology. Furthermore, the application of lead/lag time series analysis coupled with small multiples and clustering provides users with a novel mechanism to compare similarities of attributes between countries which possibly responded to the same known external disruptions in the network. This case study also explores the use of a diffusion graph to visualize the trade influence over countries, coupled with cluster analysis. Finally, the integration of correlation analysis with anomaly detection enables a novel visual analysis of the impact of trade-related anomalies on social instability, creating new mechanisms for hypothesis generation.

In order to identify challenges and needs in this area, an iterative design procedure was followed through with collaborators in Sustainability and Political Science. From the discussion with the collaborators, a basic structure of analysis was defined (detailed in Section 5.2) and system components identified. The basis of analysis is supported by an interactive map to visualize both the volume and proportion of trade events for imports, exports, and triadic structures. Temporal correlation analysis and anomaly detection are integrated to guide the user to structures of interest within the data, and small multiples are used to allow comparisons between regions with similar anomalies. Other views include a trade diffusion graph which explores the propagation effect of trade anomalies and a hierarchical clustering view which groups countries based on their triad profile similarities or trade partner similarities.

Overall, this case study advances previous visual analytics methods in this area by providing a novel correlation and triad anomaly detection procedure to reduce the analysis search space. A suite of well-known visuals were utilized and several

customized views were provided to support advanced data exploration and hypothesis generation.

## 5.1 Global Trade Analytics

The aim of this case study is to support domain experts from political science and sustainability to explore relationships between international trade structures and regional stability measures (e.g., national economic development measures and regional conflicts). In the initial discussions with our stakeholders, our goal was to elicit their analytic tasks and requirements. The requirements of the system was formatted based on the rationale of the targeted analysis from experts as well as relevant literature in trade analysis [142–144].

### 5.1.1 Data Description

To study the relationship between international trade events, country stability, and human conflicts, the domain experts have collected and provided international trade data [145], Armed Conflict Location Event Data (ACLED)[110], and economic data.

**Trade Data:** The trade data comes from the United Nations Food and Agriculture Organization (FAO) [145], and contains worldwide bilateral trade information of agriculture products. Each record in the data represents the total volume of trade occurred between two countries within one calendar year, and each record has 6 attributes: year, exporting country, importing country, the product of trade, trade quantity in tons, and trade value in dollars. The global trade data is available from FAOStat [145] and has been aggregated yearly from 1986 to 2013. The original data contains 600 types of agriculture products, many of them are very similar, such as “Meat, pig” and “Meat, pork”. Our collaborators have grouped the trade products

into 20 categories creating a two-level hierarchy. As such, the trade network data consists of approximately 8.8 million trade records over 245 countries and territories with 20 product categories made up of approximately 600 individual goods. This means that there are nearly 150,000 possible networks when considering countries, product categories and individual goods (even more for group combination selections).

**Stability Data:** The Worldwide Governance Indicators (WGI) dataset [146] is used, which reports aggregate governance indicators for over 200 countries and territories from 1996 to 2015. The original data contains a six-dimensional governance index. Our collaborators focus on the “Political Stability and Absence of Violence/Terrorism” governance index which ranges approximately from -2.5(weak) to 2.5(strong); high governance index indicates high stability.

**Economic Data:** The GDP data from the world bank is also collected. It contains GDP and GDP per capita data for 265 political entities from 1960 to 2015, some of the political entities are continents. Only the data that overlaps with countries in the FAO data is used.

### 5.1.2 Design Requirements

When beginning the system development, the analysts from sustainability and political science identified the primary question that they wanted to ask of their data: “Are there topological network structures associated with trade that serve as potential indicators of future instability?” Based on our discussions, it is clear that this topic has received much attention from their communities; however, our domain experts had identified current limitations in studying global trade networks. Specifically, many studies that do attempt to study the topology of food trade networks suffer from three typical simplifications that, while making analysis more tractable, may significantly distort conclusions:

- Aggregating commodities into a single network of “trade” (e.g. [124, 147]) - this ignores the flow of individual commodities and the relative importance of each.
- Ignoring the intensity, or weight, of trade (e.g. [121, 147]) - these studies consider only whether trade exists between country A and B, but disregard the magnitude of that trade. This ignores power differentials between countries, economies of scale, and concentrations of resources in specific regions of the network.
- Ignoring the directionality of the trade (e.g. [127, 147–149]) - perhaps the most egregious simplification is that an importer and exporter are treated equally and the direction of the flow of goods, transfer of risks, and shifts in power are ignored.

As such, our experts wanted a system that could allow for both aggregated and disaggregated food networks. We also discussed the fact that the combination of goods and years results in over 150,000 networks and this number is growing over time. What they required was a tool that could highlight time periods where large changes in the network structure were occurring and when these changes were correlated with changes in the various stability or economic measures that they had collected (e.g., ACLED events, stability indicators, and GDP). As such, our design rationale follows the visual analytics mantra proposed by Keim et al. [150], “Analyze first, show the important, zoom, filter and analyze further, details on demand,” and we have formulated several design requirements for the system. These design requirements also reflect the components of the proposed visual analytics framework in Chapter 3.

**R1** The system needs to provide an overview for the expert to start their analysis given the sheer number of possible networks, each of them can be centered on many possible countries. This overview needs to visualize patterns detected

from the networks and the possible relationships between the network and local instability to help the expert select the data needed to be explored (Ch. 3: **C1**).

**R2** Explorations of a trade network should be supported for the investigation of different trade measures, e.g., imports, exports, and reliance, and the structure of country's local network topologies (Ch. 3: **C1**).

**R3** The system needs the capability to show the possible influence between countries on their trade network (Ch. 3: **C4**). This is important to analyze the impact of food shortage on the network, and how such impact could further link to a country's stability disruption.

**R4** Statistical analysis, such as correlation analysis and anomaly detection, should be integrated to assist the expert in generating and verifying hypotheses (Ch. 3: **C3**).

**R5** The system needs to retrieve similar countries based on the selected anomaly, and allow the expert to compare the trade networks involved with these countries (Ch. 3: **C2**, **C3**). Supporting this is critical for the expert to generalize their findings.

In this problem domain, we have analysts that need views of space, time, and network connections of entities. Spatial views need to allow quick reference to regional variations in form analysts are familiar with (e.g., country level choropleth maps), while the temporal components need to be designed to help explore lags and leads. Critical to this is identifying regions that have similar temporal trends to a country of interest. If known disruptions have occurred in country A, analysts want to explore if other countries are experiencing similar trends as this may indicate upcoming disruptions (e.g., social unrest or violence). This suggests the needs for multiple views designed to link spatial regions for selection, highlight common attributes amongst regions for



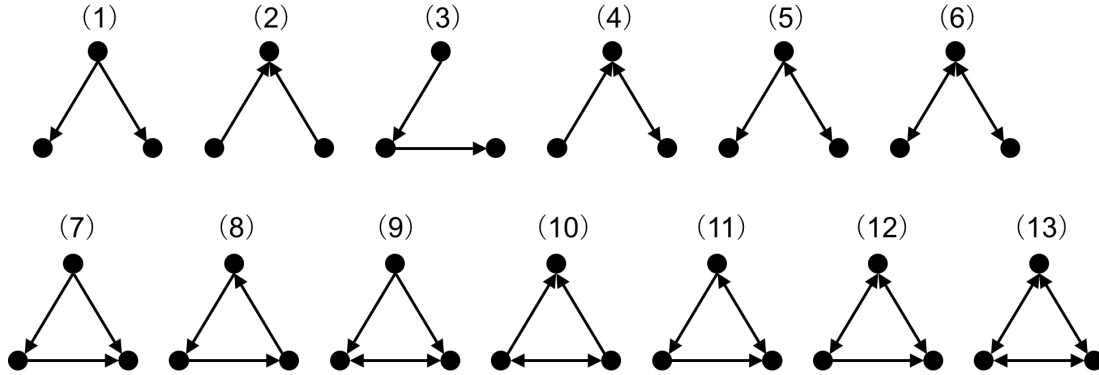


Figure 5.2: The 13 Possible Triad Configurations Are Labeled with Numbers in Braces. In the First Row, Triad 1 to 6 Are Open Form Triads, While in the Second Row, Triad 7 to 13 Are Closed Form Triads.

clustering, and identify unexplored regions that exhibit similar temporal and network dynamics that may have been leading indicators of unrest in known problem areas.

### 5.1.3 Triadic Analysis

Our domain experts also had specific requirements for analyzing local network attributes. Specifically, they wanted to explore triadic structures of the global and local trade patterns. Characteristics of trade networks, as with other types of networks such as social networks, biological networks, and economic networks, can be revealed by analyzing local structures. For example, if country A imports a specific product from both country B and country C, a competitive relationship may arise between B and C. The nature of this relationship also depends on the product flows between B and C and the types of products flowing between all three countries. To analyze these effects of local structure in a directed network, researchers typically analyze the network's triads, or subsets of three connected nodes. In contrast to dyads, triads are the smallest structures within a network that exhibit truly social characteristics [151–153]. The use of triads has long been a standard methodology in

the study of networks [152], particularly with regard to properties such as structural balance [154] and transitivity [155]. While larger subgraphs can also be studied (e.g. tetrads), triads are a logical starting point for directed networks because the relatively small number of possible configurations are manageable for exploratory analysis [143].

A triad is a three-node directed subgraph, of which there are 13 configurations (grouped triads) (Figure 5.2). When a network represents international trade, a triad can be viewed as a group of three countries and their associated trade interactions. The reasons we use triads, instead of dyads (two-node relationship) are: 1) Dyads result in fewer trade patterns when compared to triads, and therefore using triads may provide a finer characterization of the network's local pattern; 2) Bonds coincide more with triads than dyads, and; 3) Triads can be considered as more stable than dyads because a triad describes a relationship among three countries which is less easy to change [156].

Triadic analysis has revealed previously unknown structural attributes of networks that appear to be fundamental to a wide variety of natural, social, and man-made systems [143, 157–160]. For instance, one of our collaborator's previous studies shows that countries at different stages of economic development have different triadic profiles [142]. As such, our collaborators were interested in analyzing how these triadic patterns within trade networks affect a country's vulnerability. Collapsing a network into a triad profile enables users to compare across a variety of network sizes and types and to identify triads that occur more or less frequently than expected [143, 144]. In addition, each triad has particular properties (e.g., transitivity) associated with conflict and reciprocity in social interactions [155]. These properties can help explain local differences and global patterns in networks which map human social interactions [158, 161, 162].

Several studies have used triadic analysis to study relationships between countries [123, 163–165]. However, while researchers generally agree that triadic analysis is a crucial methodological component of analyzing international relationships, there is no general agreement on what triads tell us about those relationships. To better understand the value that triadic analysis offers for the study of the global network, and to generate meaningful hypotheses, there is a need for a generalized platform to explore the triadic structure of international networks, such as trade networks, and global outcomes, such as country-level stability. Thus, this system integrates triads to support trade network analysis. Grouped by triad selection, this system supports anomaly detection, correlation analysis, clustering and similarity comparisons.

#### 5.1.4 *Dependency and Leverage*

Besides measuring the absolute value of trade quantity/values from one country to the other, our collaborators also required a means of exploring dependency and leverage between countries. This measure is calculated based on the percentage of the quantity/values of the trade relationship over the total import quantity/values of the importing countries. More precisely, let  $c_i$  denote the exporting country and  $c_j$  the importing country, and the amount of exports from  $c_i$  to  $c_j$  denoted as  $Q_{c_i \rightarrow c_j}$ , then the percentage of export from  $c_i$  to  $c_j$  over all imports to  $c_j$  can be calculated as:

$$I_{c_i \rightarrow c_j} = \frac{Q_{c_i \rightarrow c_j}}{\sum_k Q_{c_k \rightarrow c_j}}$$

A large  $I_{c_i \rightarrow c_j}$  indicates that  $c_j$  is dependent on  $c_i$ , conversely, it also means  $c_i$  holds a leverage on  $c_j$  and in this case  $c_j$  is susceptible to having a short supply of the trade products when  $c_i$  stops export to  $c_j$ .

### 5.1.5 Clustering Coefficient

Clustering coefficient is a measure of the strength of connection in the subgraph consisted by the node's neighbors. It is therefore used in the framework to assess the robustness of the trade relations for each country. The clustering coefficient for a country measures how strongly connected its trade partners are. The clustering coefficient is traditionally defined as:

$$C_i(W) = \frac{\sum_{j \neq i, k \neq (i,j)} a_{ij} a_{jk} a_{ki}}{d_i(d_i - 1)},$$

where  $a_{ij} \in \{0, 1\}$  denotes whether an edge exists between node  $i$  and node  $j$ , and  $d_i$  represents the degree of node  $i$ . The edge between  $i$  and  $j$  can be in either directions. The traditional clustering coefficient does not factor in the weight of the network edges (as was noted in the design requirements), thus our system also implements an extension to the clustering coefficient as proposed by Onnela, et al.[166]:

$$\tilde{C}_i(W) = \frac{\sum_{j \neq i, k \neq (i,j)} (\tilde{w}_{ij} \tilde{w}_{jk} \tilde{w}_{ki})^{1/3}}{d_i(d_i - 1)},$$

where  $\tilde{w}_{ij}$  represents the weight of edge between node  $i$  and node  $j$ . This definition extends the traditional clustering coefficient definition to deal with weighted networks. Thus, for a given product or category, we can measure the clustering coefficient by year and detect anomalies and calculate correlations to other data.

## 5.2 System Design

Based upon the design requirements and analytical needs discussed in Section 5.1.2 and Chapter 3, a visual analytics system is developed to facilitate the exploration of relationships between international trade data and measures of a countrys stability. This system (Figure 5.1, Figure 5.3) consists of 5 major components designed to support the design requirements (R1-R5). 1) The country-product matrix (Figure 5.3)

which highlights the significant correlations and anomalies for each country/item pair (**R1**, **R4**). 2) The time series view (Figure 5.1(1)) that displays the trend and anomalies of trades and stability measures for the selected country (**R4**). 3) The choropleth map view (Figure 5.1(2)) that displays the trade profile of a selected country (**R2**), along with a small multiple of choropleth maps (Figure 5.1(3)) used to compare the trade profiles of related countries to the selected one (**R5**). 4) The clustering view (Figure 5.1(4)) and scatter plot view (Figure 5.8(3)) are used to explore the similarities and differences of each country's trade profiles and compare trade profiles to stability measures (**R5**). 5) The trade diffusion graph (Figure 5.1(5)) displays the trade flow of the selected food products from the selected country to the countries whose disruption in trade are affected or affecting the selected country (**R3**).

A sample analytic flow can be described as: 1) The analyst uses the matrix view to find the countries and food products whose trade profile exhibits a temporal pattern that is strongly correlated with the stability measures. 2) Given the selection, the time series view will display the trend of trades along with the trend of stability measures and mark the anomalies. The analyst can select a time range by brushing on the time series or clicking on an anomaly. 3) After a time range is selected, the analyst uses the choropleth map to examine the trade relationships between the selected country and other countries and uses the small multiple of maps to compare the country of interest to the trade profiles of countries that exhibits the similar anomalous pattern. 4) The analyst then uses the network diagram to explore and hypothesize on the propagating effect of trade disruptions from/to the selected countries. 5) Finally, using the clustering view and scatter plot view, the analyst explores how the selected country's trade profile compares among other countries and examines whether countries with similar trade profiles also share similar social stability measures.

This sample analytic flow highlights the basic structure of this system. Here, each view and analysis process was chosen to support specific tasks of the analysis, where the support of the coupled interactions of space, time, and network analyses is needed. By supporting multiple linked views, the switch between these three concepts can be facilitated. Furthermore, by providing methods of clustering and filtering across measures in space, time and networks, this system enables the integration of these concepts for advanced analysis.

### 5.2.1 *Correlation-based Country-Product Matrix*

Due to the large number of possible networks, the system provides a country-product matrix view to browse for interesting networks. In this view (Figure 5.3), countries and products are presented as rows and columns. Each cell in this matrix visualizes the correlations between trade measures and stability measures given a country and a product.

This matrix view consists an abstract view (Figure 5.3A) and a detail view (Figure 5.3B). The abstract view provides an overview for the temporal correlations between each country’s trade profiles and its stability measures with respect to 20 food categories. The color of the cell represents the total number of correlations, the darker, the higher. To help identify countries/product groups with large numbers of correlations, the matrix, by default, is ordered using a 2D sort method [167] to shift darker cells to the top left corner. The analyst can browse the matrix using the country slider on the left and the product slider on the top. A snapshot of the whole matrix is displayed on the top left corner to help keep track of the current location within the matrix, which is highlighted using a purple border in the snapshot. Customized reordering is also supported. The analyst can click on the “row” or “col” button on the top-right corner to order rows or columns based on the sum of correla-

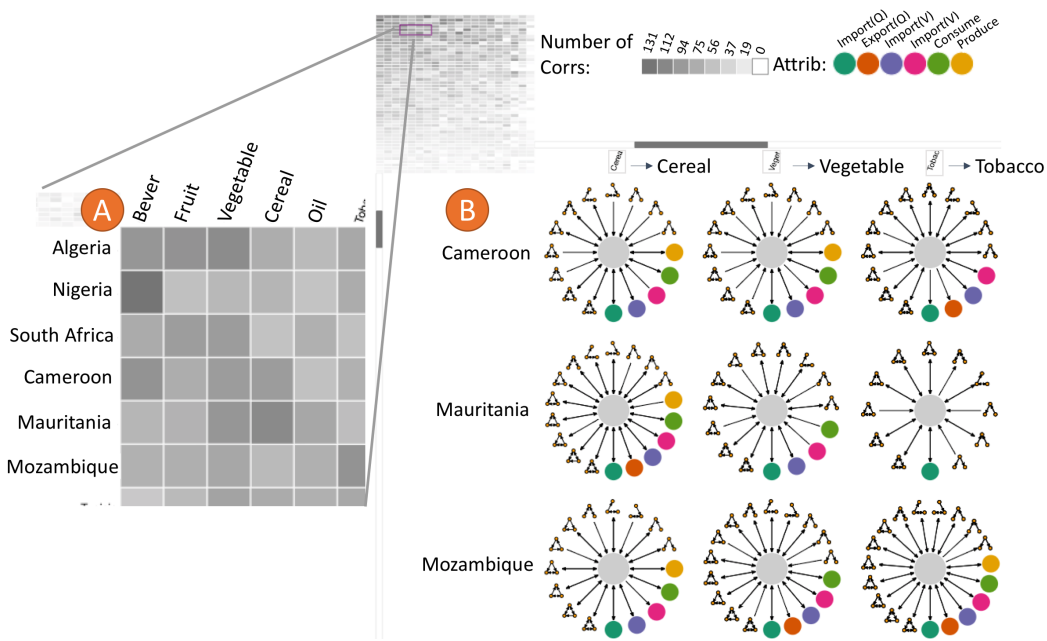


Figure 5.3: The Country-Product Matrix view Has an Abstract View (A) and a Detail View (B) to Show the Correlations Between Countries' Trade and Stability Measures. This Figure Shows Cameroon and Mauritania Have Many Correlations Between Cereal Trade Measures and ACLED Events.

tions for each row or column. S/he can also click on a particular row (country) to sort columns (products) based on the number of correlations associated to this country and sort the countries based on their similarities of the indicated correlation patterns to this country. In addition, S/he can click on on a particular column (product) to sort countries based on the number of correlations in this product.

Once interesting cells are found, the user can brush over the cells to zoom in and switch to the detail view, which displays individual correlations. For any pair of trade metric and stability, the pairwise Pearson's correlation coefficients were calculated for leads/lags between -5 and 5. In the detail view, each cell is expanded with more information. The trade attributes with correlations are displayed in a circular layout

surrounding the stability measure (colored in gray). The arrows connecting them indicate the directions of the correlations: if the correlation occurs when the trade attribute lags the stability measure, the arrow points from the trade attribute to the stability measure. Conversely, if the stability lags the trade attribute, the arrow points to the trade attribute. Double-direction arrows indicate lags in both directions.

These correlations only provide a hint of possible connections between the trade and stability for the countries. Some identified correlations may be spurious. However, these correlations serve as a starting point for the analysts to choose an initial country/item when there is no clear target in mind (knowing which particular country and food item to look at). Once the analysts pick a country/item pair, they can click on the corresponding cell to further investigate. Here we note that correlation is not equal to causation, so the method of identifying all correlations may suffer from p-fishing. As such, a human in the loop is critical for identifying potential correlations of interest and removing obviously spurious correlations.

### 5.2.2 *Anomaly Time Series*

While correlations are critical for hypothesis generation, our analysts also wanted to be cued to anomalous changes in the trade network or stability time series. By finding anomalies, domain experts can begin reasoning about what other world events may have been contributing to these changes. To support such analyses, an anomaly time series is implemented for the selected country and visualized detected anomalies.

An anomaly is defined as any sudden rise or drop from the previous year, which means there could be a shock in the agriculture trade network. To detect anomalies, the time series is first detrended by taking the first difference:  $\tilde{y}_t = y_t - y_{t-1}$  for  $t \in [2, T]$ . Then the mean  $\mu$  and standard deviation  $\sigma$  of  $\{\tilde{y}_t | t \in [2, T]\}$  are computed and the upper and lower limit is defined as  $UL = \mu + 2\sigma$ ,  $LL = \mu - 2\sigma$ . Thus, for



any time  $t$ , a sudden rise ( $\tilde{y}_t > UL$ ) or drop ( $\tilde{y}_t < LL$ ) are detected as an anomaly. Other methods, (e.g. ARIMA [168], EWMA [169]) could also be applied.

For a selected country/item pair, the anomaly time series (Figure 5.1 (1)) displays its trade measures, counts of different triads, and social stability (ACLED count and stability) by year. Which type to display can be toggled using the left-side buttons. In addition, the triad counts can be filtered based on the magnitude of the trade relationships using a slider in the top of the interface. This filter is applied to all the triad counts in the networks under analysis throughout the entire framework. For all time series, anomaly detection has been conducted and detected anomalies will be plotted as dots on the corresponding lines. Each anomaly dot can be clicked to investigate countries sharing the same type of anomaly (which occurs in the same year of the same attribute time series). These countries will be displayed as small multiples.

Since there can be multiple time series in the line chart, the analysts can click on the legend to highlight or de-highlight a time series. Whenever the time series is highlighted, a temporal correlation indicator will be displayed to show the correlation between the trade time series and the stability time series from lag 5 to lead 5. The correlations are colored using a diverging color scheme in which the blues encode positive correlations and reds encodes negative correlations. These correlation bars are also shown under each small multiple map to indicate the corresponding correlations for that country.

### 5.2.3 *Choropleth View*

For a selected country and product, we have a main map and a set of small multiples to visualize trade network elements.

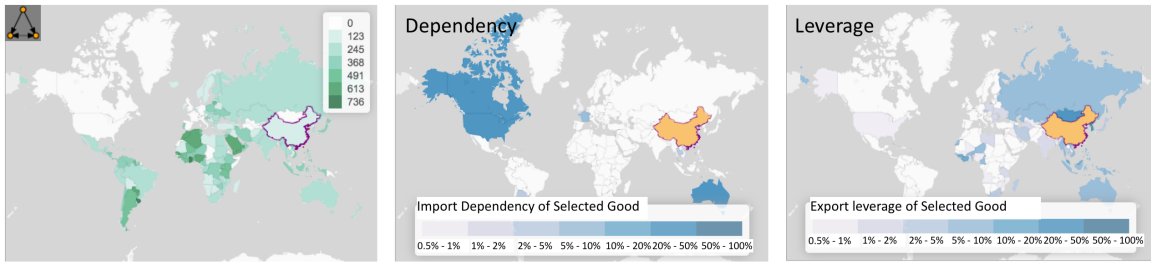


Figure 5.4: The Main Choropleth Map Can Be Used to Visualize Triad, Trade Quantity/Value, and Country Dependency/Leverage. Examples Show the Triad Distribution (Left), China’s Dependency (Middle), and China’s Leverage (Right) in Cereal Network.

**The Main Choropleth Map.** The main map (Figure 5.4) interactively displays the global relationship of other countries to the selected one. It can be switched between trade measures, dependency and leverage, and triad distributions. Users are also allowed to quickly change the selected country by clicking on a country on the map. This choropleth map also supports a detailed overview of a country’s core trade products. When mousing over a country, the top 5 import/export products for the country are shown as a bar chart. The value in each bar represents the ratio of the goods’ import/export value (or quantity) over the country’s total import/export value (or quantity).

*Trade Connections and Magnitude.* On the main map, the selected country is highlighted in orange and all other countries on the map are colored using a sequential color scale based on the selected measure (trade quantity/value or triad). **Triads** are visualized by the sum of the triad counts in the selected types (out of the 13 types) associated with each country in the food network. **Trade Quantity** is given in tons. When this option is chosen, countries that are in the selected country’s trade network are colored based on the imports/exports quantity from/to the selected country. A

logarithmic scale is used for better color separation. **Trade Value** is given in dollars. Again, when this option is chosen, countries that are first order members of the selected country's trade network (which means they are directly trading with each other) are colored based on the dollar value of imports/exports from/to the selected country. This color also employs a log scale.

*Trade Dependency and Trade Leverage.* **Trade dependency** is derived from trade quantity and trade value. The map can show the trade dependency of the selected country to the other countries. The dependency, as explained in Sec. 5.1.4, has a range of  $[0, 1]$ . The color of each country represents how dependent the selected country is to the country. The dependency can be toggled between import and export. The color scale is manually defined to be  $[0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1]$  which is approximately logarithmic. **Trade Leverage** is also derived from trade quantity and trade value. The map shows the trade leverage of the selected country to the other countries as described in Section 5.1.4. The colors are defined in the same way as the dependency.

**Small Multiple Maps.** To help the analyst find similar countries in terms of anomalies, this system has a small multiple maps view (Figure 5.1 (3)) alongside the main map such that for a selected anomaly, the small multiple maps view will display all other countries that have an anomaly in the same year with the same attribute. These small multiples also use choropleth maps to visualize the trade measure centered on these similar countries. The visual setting is synced with the main map. Each of these small multiples was coupled with a colored correlation bar to indicate the temporal correlation of the selected trade attribute to the stability measure. These correlation bars have the same visual encodings as the one in the anomaly time series. Clicking on the label of any small multiple maps will switch the centered country between the

main map and the small multiple map, and the other components of the system will update according to the new country selection.

#### 5.2.4 Trade Diffusion Graph

The trade diffusion graph (Figure 5.1 (5)) is used to display the potential impact of an anomaly in the trade network. When there is a sharp decrease or increase in trade for a particular food product, it is tempting to think how such change could be caused by or be causing the trade disruptions of other countries. For example, when there is a sharp decrease in wheat export from China in 2004, many countries experienced a sharp decrease in wheat import in the same year, analysts are interested in knowing whether these countries are directly or indirectly affected by the export drop of China. To explore such questions, we developed the trade diffusion graph, which displays a directed acyclic graph (DAG) connecting the influencing countries to the influenced countries in the same year. Building this graph requires two steps.

*Step 1: Identifying source/target countries.* Once an anomaly is selected, the list of corresponding trade diffusion source/target countries is queried according to the following rules:

- If the selected anomaly is a sharp decrease (or increase) in exports, the target countries are identified with a sharp decrease (or increase) in imports in the same year.
- If the selected anomaly is a sharp decrease (or increase) in imports, the source countries are identified with a sharp decrease (or increase) in exports in the same year.

The countries with sharp changes in exports are considered to be the sources of the trade diffusion, while the countries with sharp changes in imports are considered to

be the targets. The identified graph will have only one source/target (the selected country) but multiple targets/sources (the identified countries).

*Step 2: Identifying diffusion paths.* A diffusion path, in our scenario, is defined as a trade path from the source to the target with all links (export to import) that have a higher dependency than a user-defined threshold. To determine if there is such a link from country  $A$  to country  $B$ : country  $A$  may influence country  $B$  when  $I_{A \rightarrow B} > \epsilon$ , where  $\epsilon \in [0, 1]$  is a user-defined threshold. The depth-first search is used to retrieve these paths, and prune to only keep the paths from the source(s) to the target(s). When the selected country is a source, a DAG starting from the selection is obtained, and when the selected country is a target, a DAG ending to the selection is obtained. The graph is drawn using Dagre Javascript library and Jünger and Mutzel’s crossing minimization method [170], and the layout of the graph positions the countries based on their topological orders to emphasize the flow of influences from the source country(s)(left) to the target country(s)(right). The strength of dependencies are encoded using the width of edges, and each node is labeled by the flag of the corresponding country. Mousing over each country will highlight all the paths between this country and the selected country. The selected country is colored in orange and the other anomalous countries are colored in green. A slider is also provided for interactive adjustment of the dependency threshold.

### 5.2.5 Clustering and Comparison

Our analysts wanted the ability to quickly explore similarities across trade and triad profiles and cluster countries based on these attributes. They would like to see if similar countries might have the same vulnerability in the trade network or may be impacted by the same change happened in the network that could cause some social instability.

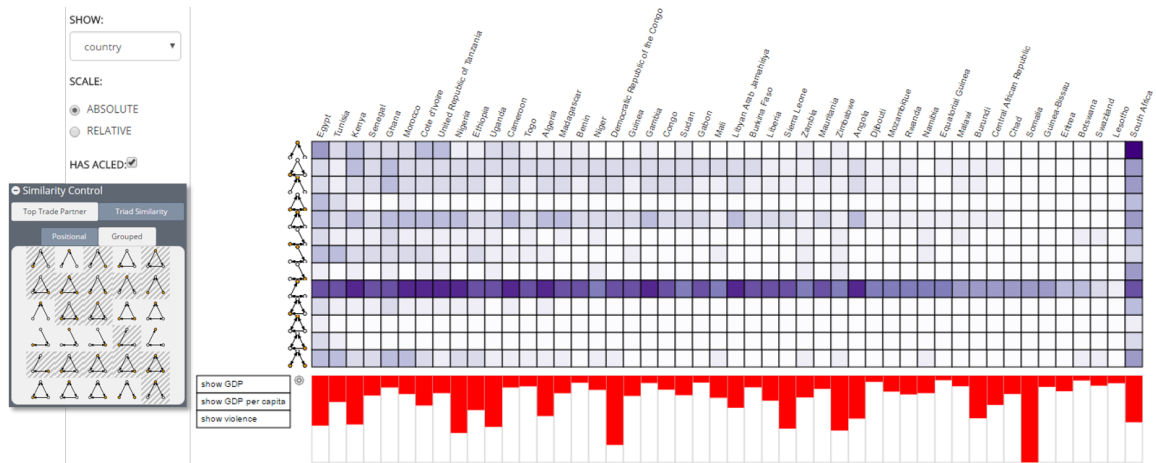


Figure 5.5: The Matrix View Showing the Triad Distribution for Each Country. Rows in the Matrix Represent Triad Configurations and Columns in the Matrix Represent Countries. Each Cell in the Matrix Represents the Triad Count for The Country. The First Column in the Matrix Is the Currently Selected Country in the Map View. The Other Columns Are Ordered Based on Their Vector Similarity to the First Column. The Analyst Can Remove Some Triads (Rows) from the Matrix by Clicking on the Triad on the Similarity Control Panel. Doing So Will Recompute the Vector Similarities and Reorder the Matrix. The Bar Chart in the Bottom Shows the Instability Measure (e.g. ACLED, GDP).

**Scatter Plot.** To enable further comparison and correlation analysis, a scatter plot view (Figure 5.8 (3)) has been created displays the countries' network attributes versus ACLED counts, GDP, or GDP per capita. The analysts can use the selected boxes next to each of the axes to toggle either the attributes or scales (linear, square root, and log).

**Clustering Visualization.** This system has integrated a hierarchical clustering view for interactive clustering analysis. The countries are clustered using a complete-link hierarchical clustering algorithm. If the hierarchy is cut using a similarity threshold,

the nodes whose minimum similarities are at least equal to the similarity threshold can form clusters. The clusters are displayed in a dendrogram (Fig.5.6). In this view, the clusters are represented by a box enclosing a set of rectangles, where each rectangle represents a country. The internal nodes of the hierarchy are drawn above the boxes with a similarity threshold at which the descendants will merge. The number next to each box represents the minimum similarity of countries within the box, and the number next to each internal node represents the minimum similarity of all countries under this node. Initially, the clusters are formed according to a default threshold, 0.5. The analyst can drag the slider to change the similarity threshold or click on a box or an internal node to expand or collapse the cluster. Clicking on any box will expand the hierarchy by splitting the countries in this box into two clusters and make the current box an internal node connecting these two clusters. Clicking on an internal node will collapse the hierarchy by joining all the boxes below the node into a larger box, and the internal node will be replaced by the newly created box. The color in each of the internal rectangles encodes the value of either the GDP, the GDP per capita, or the number of violent events in the country which the rectangle represents. Clustering utilizes partner similarity or triad similarity.

*Triad Similarity.* This metric defines the country similarity by comparing their triad profiles. A country's triad profile is defined as a vector of the frequencies that the country appears in each triad configuration. Thus, when using positional triad configurations, the country's triad profile is a  $13 \times 1$  vector where each entry represents one frequency. The triad similarity of two countries is then defined as the vector similarity of their triad profiles. We used Euclidean distance to calculate this similarity. By default, all entries in the triad profile are used for the similarity calculation. The analyst can filter out some types of the triad configurations using the control panel

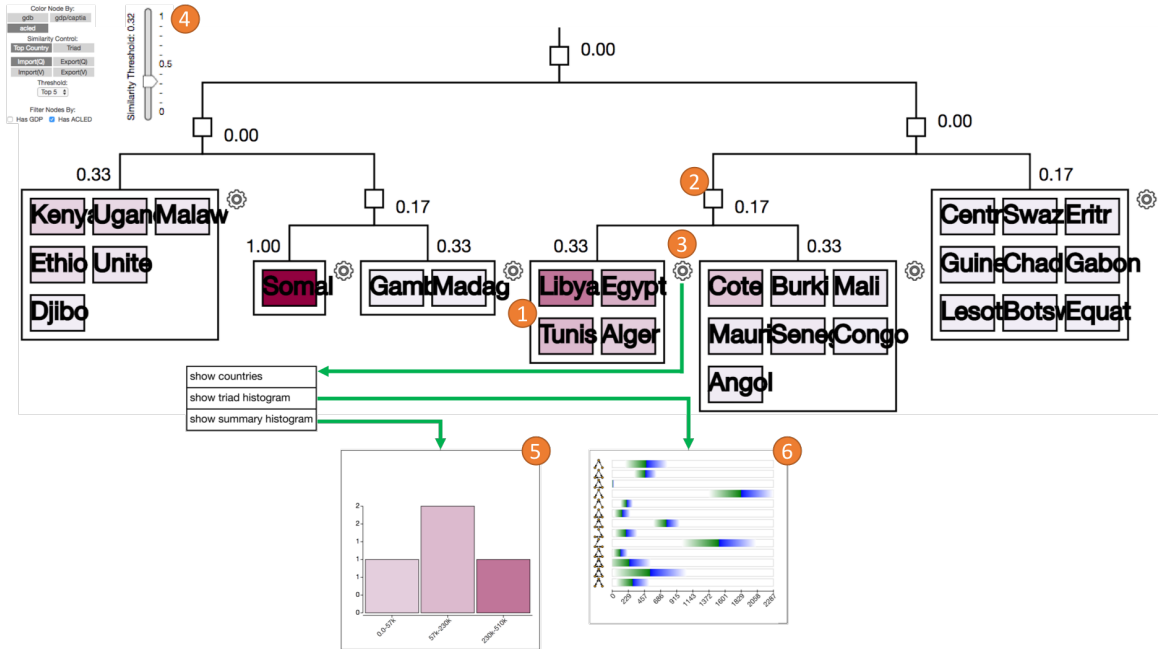


Figure 5.6: The Clustering View Displays the Clusters of Countries in a Dendrogram. Countries in a Cluster Are Grouped in a Box (1). Clicking on Any Box Will Break the Cluster into Two Smaller Clusters. Clicking on Any Internal Node (2) Collapses All Children. The Analyst Can Use the Slider (4) to Adjust the Similarity Threshold for the Hierarchical Clustering. The Color of the Countries Encodes Their Selected Stability Measure. By Clicking on the Setting Icon (3) at the Top Right Corner of Each Box, the Analyst Can Choose Between Different View Options for the Fox. The Box Can Change to a Histogram of the Stability Measure (5) or a Bar Chart That Compares the Mean and the Standard Deviation of Every Triad Configuration (6).

(under the similarity control area) so that the similarity is calculated only on the selected entries.

*Trade Partner Similarity.* This metric defines the country similarity by comparing its most important trading partners in the network. Given country  $c_i$  and  $c_j$ , we can define the set of their trade partners as  $T_{c_i}$  and  $T_{c_j}$ , and then the partner similarity



between these two countries can be defined as  $|T_{c_i} \cap T_{c_j}|$ . Interactive filters can be employed such that  $T_{c_i}$  can represent only the  $N$  largest trade partners. In this way, we can look for countries whose largest trade partners share a large set overlap.

**Cluster Exploration.** Detailed exploration of the grouped country in a box is available in this system. In the top-right corner of each, a configuration icon is found. The analyst can click on the configuration icon to switch between three visualization options for displaying the country cluster. The default is to list the countries, as displayed on the top of Figure 5.6. The second option switches this view to a histogram, as shown on the lower left of Figure 5.6. It can be either the triad summary histogram or the data summary histogram. In the data summary histogram, the violence count, GDP or GDP per-capita of all countries within the cluster are summarized. In the triad summary histogram, as shown in the lower right of Figure 5.6, each bar shows the mean of the count of one type of triad configuration for all countries in the cluster. The length of the green color area represents one standard deviation below the mean, and the blue color area represents one standard deviation above the mean.

**Triad-Country Matrix View** In order to quickly compare triad profiles by country, the triad profiles of each country are visualized in a matrix view (Figure 5.5). Each column in the matrix represents a country and each row represents one triad configuration. Each cell in the matrix is colored based on the magnitude of the frequency the country appears in the corresponding triad configuration. The color scale can be toggled to range over the entire matrix (absolute) or be defined separately in each column (relative). The first column (country) in the matrix is determined based on the current selection of country in the map view, and the rest of the columns are ordered based on the similarity between their triad vector and the triad vector in the first column. The analyst can click on the similarity control panel to toggle between positional triad profiles and grouped triad profiles. The analysts can also click on any

triad figure on the similarity control panel to remove (or add back) the triad row from (or into) the matrix, and the order of the columns will be recomputed each time the analyst does so. The bottom of the matrix view shows the magnitude of the chosen instability measure. The matrix can also be toggled to view the triad profiles for the trade products instead of countries.

### 5.3 Usage Scenarios

Our usage scenarios were developed by our domain experts in political science and sustainability. Training and paired analysis was used initially, and then experts took our tool and integrated it into their current data analysis process. In this section, the details on their findings are presented, and hypotheses generated when using this visual analytics system are discussed.

#### 5.3.1 2011 East Asia Drought and Flood

Thailand is a major rice exporter globally and is typically subject to severe flooding. However, Thailand has recently suffered several years of extreme drought beginning in 2011 [171]. Our analysts were interested in performing an exploratory analysis focused on the global impacts of localized climate events. In our first case study, we demonstrate how this tool was used to develop hypotheses about how changes in Thailand's rice exports might induce potential social unrest in Africa.

First, the analyst selected rice and Thailand. The line chart showed a significant decrease in the quantity of rice exports in 2012 (Figure 5.1(1)). While this was not surprising given Thailand's agriculture disruption in 2011, it did establish that the drought was followed by a drop in exports the following year. To see which countries might be significantly impacted due to this export drop, the analyst clicked on the anomaly dot on Thailand's export line. Small multiple maps listed all countries

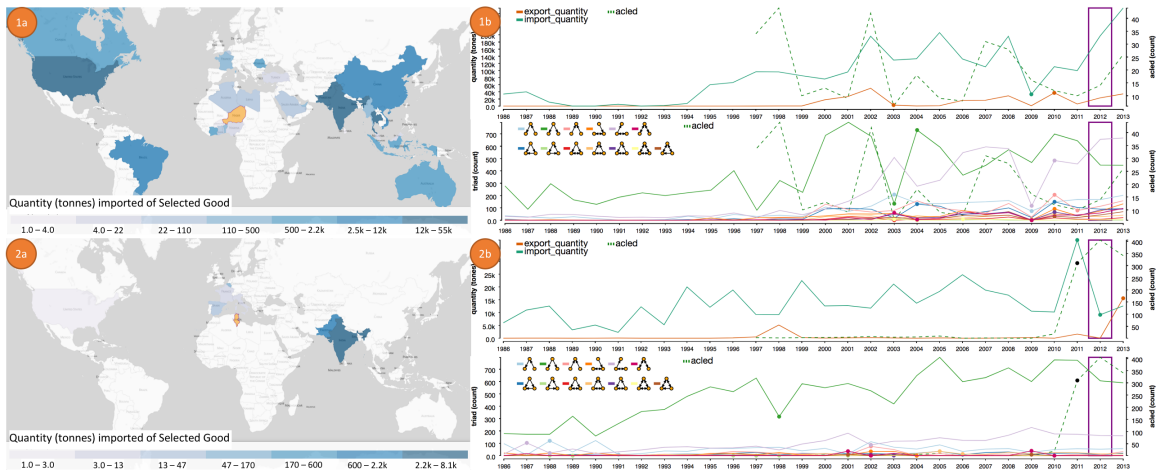


Figure 5.7: Comparison Between Niger (1) and Tunisia (2) on Each Country's Import Quantity of Rice. By Comparing Both Countries' Major Exporters of Rice Using the Choropleth Views (1a, 2a), It Can Be Seen that Niger Imports from a Wide Variety of Countries Whereas Tunisia Imports Primarily from India and Pakistan. The Time Series for Both Countries (1b, 2b) Show That Tunisia Had an Import Quantity Dip in 2012 While Niger Had No Such Disruption. At the Same Time, Tunisia Had a Sharp Increase in Local Conflict Events. The Analyst Can Also Observe That Tunisia Had Fewer Types of Triads Than Niger.

which also had significant drops in rice exports in 2012. On the trade diffusion graph (Figure 5.1(5)), countries that experienced a decrease in rice imports were retrieved and linked by their dependencies to Thailand. The resulting graph showed that some countries, such as Niger, United States, India, China, and Vietnam, transferred this impact to other trade partners but were not significantly impacted by Thailand's rice export shrinkage, while many other countries, marked with the green background, suffered large decreases in import quantities of rice.

The analyst then wanted to explore why some countries suffered from Thailand's rice export decrease more significantly than others. In particular, he wished to know

if this might be related to their trade network structures. He clustered these countries shown on the trade diffusion graph using triad similarity. The clustering result, with a threshold of 0.5, is shown in Figure 5.1(4). A finding of interest to the analyst is that the largest group (Tunisia, Niger, etc.) contains most of the countries that were affected by Thailand's drop in rice exports. When the analyst split the largest grouping, he noticed that the remaining countries in the group were all the countries that experienced the rice anomaly, except Fiji. By switching to view the average triad profile in each cluster, the analyst noticed most of the variability among these clusters came from the open triads (type 1 to 6 on Figure 5.2). Thus, the analyst then chose to remove the closed triads in the clustering control panel and re-cluster the countries. He found that the countries with import anomalies were more tightly grouped together. Looking at the trade diffusion graph again, the analyst quickly noticed that countries that were initially out of the largest cluster had leverages on multiple countries. For example, Pakistan, though marked with the green background, also served as a transferring country in the network as it had multiple leverages on Iran, Kuwait, and Tunisia, and it had a large amount of the Figure 5.2 type (2) open triad. All the other anomalous countries had no leverages on other countries, and Fiji, along with most other countries that were later separated from the largest group (Australia, Cambodia, and Niger) only leveraged on one country in this graph, and this group of countries had significantly less type (2) triads. These findings indicate that by using triad profiles, it may be possible to group countries based on their leverage on other countries. Our analyst noted that this can be a testable hypothesis for future research.

Recalling that our analyst was interested in finding possible relationships between a country's trade network and its social stability, he decided to further explore Tunisia and Niger which, in general, have local conflict problems. He sought to understand

why Tunisia had an anomaly while Niger did not, and whether this anomaly could be relevant to their local unrest events. The analyst clicked on these two countries to check their import quantity and ACLED event line charts. As shown in Figure 5.7, Tunisia had an import quantity dip in 2012 while Niger had no such disruption. At the same time, Tunisia had a dramatically increased number of local conflict events, and the ACLED line shows a sharp rise during 2011 and 2012. Although the exact relationship between rice imports and local armed conflicts cannot conclusively be determined here, it nevertheless provides evidence that these two variables may be related, as their anomalies happened close in time.

To see how this is related to Niger's and Tunisia's rice trade, the analyst inspected their worldwide import quantities. As shown in Figure 5.7, Tunisia imports most of its rice from India and Pakistan, both of which are in the trade diffusion graph of Thailand's export cut. However, Niger imports rice from a wide number of countries, including Brazil, Canada, and Australia, which are not all in the diffusion graph and not shown in the small multiple maps (meaning they did not have significant drops in rice exports). As such, having a wide trade network that does not connect in a single trade diffusion path can increase a country's resilience to agriculture disruption. It may also enhance local stability, as Niger did not exhibit a significant social unrest problem like those in Tunisia. Regarding triads, the analyst compared these two countries on the scatter plot and their triad line charts. As expected, Tunisia had fewer types of triads than Niger.

### *5.3.2 Sustainability - Food Insecurity and Political Instability*

The second usage scenerio was provided by a domain expert in sustainability who focused on using the system from the perspective of stakeholders in the international development and aid community. In particular, he used the system to suggest

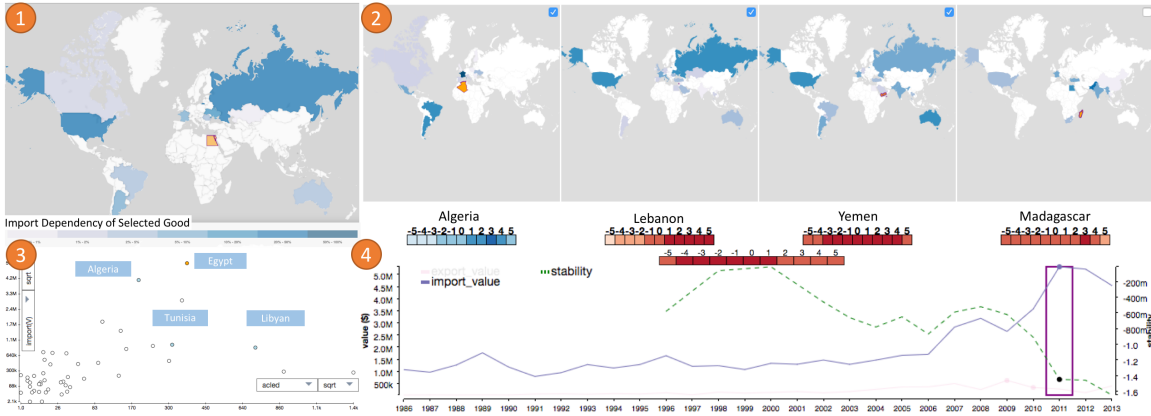


Figure 5.8: Egypt Had a Sharp Increase in Cereal Import Value in 2011. This Figure Shows (1) the Import Value Distribution of Egypt, (2) The Similar Countries Who Had the Same Increase, (3) the Scatter Plot by ACLED and Import Value of Cereal Among Africa Countries, and (4) the Time Series Indicating the Sharp Increase as Well as the Negative Correlation Between Egypt’s Import Value and Its Stability Index.

geographical regions where decision makers in his domain may want to focus their attention. In this case, the analyst sought to identify potential hotspots of political instability arising from food insecurity.

The analyst started by examining the country-product matrix and focused on cereal (a proxy for wheat in this dataset). The matrix indicated that, for both Mauritania and Cameroon, several correlations were detected between their cereal trade measures and social unrest events (Figure 5.3). Knowing that Mauritania’s main agricultural products are wheat, grain, and corn (part of the cereal category), while Cameroon’s are coffee, sugar, and tobacco, but not cereal, the analyst expected that Cameroon would rely more on imports of cereal than did Mauritania. To confirm this, he looked at the line charts for cereal import values and quantities of both countries and found that, on average, Cameroon’s import of cereal was twice that

of Mauritania. Furthermore, the analyst found that, for Cameroon, during 2011 and 2012, the import value of cereal increased significantly, though the quantity only slightly increased. This indicated a probable price increase in cereal, which matched the global wheat price hikes in early August 2011 as wheat production and markets were disrupted by natural disasters (e.g., 2010 Russian wheat crops were lost in wildfires and led to decreases its 2010/2011 production and stock). Turning specifically to wheat (a subcategory under cereal), the analyst found that Egypt (a similar country pulled from the small multiple maps) experienced a significant price increase for imported wheat. The line chart revealed that not only did Egypt's import value of cereal products exhibit high positive correlation with ACLED events, but Egypt also had an uncharacteristic increase in cereal imports in 2011 followed by an anomalous increase in ACLED events in 2013. This could suggest that the surge of ACLED events were caused partly by rising food prices.

Next, our domain expert chose to explore what other countries also experienced a sudden increase in cereal imports in 2011. He did this by clicking the anomaly indicator on the import value time series, which loaded a list of countries on the small multiple map view. The domain expert immediately noticed Algeria on the list, whose import value also had a high positive correlation with ACLED events (shown by the mini correlation bar underneath the map). Scrolling down the list, the analyst also noted Lebanon and Yemen, both of which experienced uprisings during this time. The analyst moved these countries to the top of the list by clicking on the checkbox on the top left corner of their maps and compared their cereal exporters. He noticed that all four countries imported wheat from a majority of the same exporters, with the exception that Algeria was less reliant on Russia. The domain expert then switched the target of correlation comparison to the stability index and found that Egypt, Yemen, and Lebanon all had a strong negative correlation with the stability

index, while Algeria did not. These findings suggest that the sudden increase in the import value negatively correlates with stability for Egypt, Lebanon, and Yemen and that this negative correlation might be due to the similarity of their import patterns. In addition, their triad profiles have a similar distribution and all show that triad type (5) and (1) are the most common.

Finally, our domain expert chose to explore which other countries had cereal trade structures similar to that of Egypt. His presumption was that if such countries substantially share the same global structure for wheat trade, they may be at risk for episodes of political instability similar to those witnessed in Egypt. Therefore, the analyst performed a clustering on trade partner similarity and found that Egypt, Libya, Tunisia, and Algeria were clustered together (Figure 5.6).

In assessing the veracity of these suggested focal countries, our analyst noted that in 2011, civil war erupted in both Syria and Libya, with Lebanon becoming engulfed in the Syrian conflict by 2012. Internal rifts also escalated into civil war in Yemen in 2014, while in 2013 the Council on Foreign Relations reported that the risk of political instability in Jordan had reached its highest level in over 40 years [172]. Additionally, Algeria and Tunisia experienced political unrest to varying degrees during this same time frame. While appearing to support the notion that the tool's trade similarity analysis may accurately focus a user's attention, our domain expert also cautioned that these examples are admittedly anecdotal. Both our experts found that the tool provided unique capabilities for hypothesis generation that were unavailable through their current workflow; however, the ability to couple this with analyses for hypothesis testing was further requested.



## 5.4 Discussion

In this case study, a visual analytics system for spatiotemporal trade network analysis is demonstrated. This system focuses on network feature analysis and enables users to quickly identify temporal anomalies within these features as well as correlate these features to country level stability indicators. By linking multisource data for exploration, this system enables users to explore and develop complex hypotheses. While several visual analytics systems have explored trade network analysis, this system focuses specifically on integrating triad analysis with other metrics to provide novel insights.

This system has been designed and evaluated through collaboration with domain experts from the Global Security Initiative and the School of Politics and Global Studies at Arizona State University. Examples ranged from exploring theories of Capitalist Peace to food insecurity and instability cascading through the trade network. Our partners found that the anomaly detection provided them with a clever means of identifying regions and time periods of interest, and while they relied on their own domain knowledge as a starting point, they noted that they planned to do further exploration to see what other hypotheses might be developed with a deeper dive into unexplored anomalies. Both the clustering view and the small multiples anomaly view were key features in their analytic process; however, both analysts noted that the framework could only support exploratory data analysis.

Reflecting on the design of the system, there are a variety of challenges that must still be addressed. Given the highly multivariate nature of the data, and the need to identify lags/leads in conjunction with potential network disruptions, numerous anomalies are detected. It is unlikely that the initial analysis step in the visual analytics process will be completely sufficient for discovering unexpected patterns. While

our designs attempted to provide an overview of anomalies through a country-product matrix view, new visualization designs should focus on improving the overview, as well as combining a suite of analytic methods that can vote on most likely anomalies. Here, the tradeoff of reducing the number of anomalies for exploration versus missing an event needs to be considered. The limitations of choropleth maps in our design is also noted. While these maps directly support analysts with a familiar structure, the sheer number of countries and trade goods that exist in the dataset limits their functionality. Again, these maps provide an overview and serve the analyst well for filtering; however, new views that can highlight multivariate trade over should also be explored. Sankey diagrams may serve as an alternative, or dynamic network visualization techniques (e.g. [173]) may be advantageous.

The current limitation of this system is that it still falls short in identifying spurious correlations. While the system can provide many useful clues about the interdependency of trade between countries and the effect of such interdependency on social stabilities, it does not provide a way to justify the findings and exclude other possible explanations. The analyst must rely on external sources to verify the results. The next case study explores extension of the system to support identification of spurious correlation, and new interactions and views are developed to support this task.

SPATIOTEMPORAL TRADE NETWORK ANALYSIS – A CASE FOR  
ADDRESSING SPURIOUS CORRELATION

The previous case study mainly focuses on visualizing correlations; however, little attention focused on identifying spurious correlations. Spurious correlations arise when two variables appear to be correlated given the observation but are not intrinsically related. In this case, the correlation is either a product of coincidence or due to the confounding bias (Section 2.2). Spurious correlation is a common challenge in data analysis. For example, economic data from 1875 to 1914 shows a positive correlation between tariff rate and economic growth. However, a close examination of the data reveals that the result is heavily distorted by three high tariff and high growth countries at that time: Argentina, Canada, and, to a lesser extent, the United States. The labor-scarce, land-abundant nature of these countries made their governments more reliant on tariffs for revenue, yet the economic structure of these countries led to high growth regardless of the tariffs [174]. Disregarding these underlying facts in the data can lead to an unreliable conclusion on the relationship between the tariff rate and economic growth, which may generate more erroneous results if extrapolated to other scenarios. The difficulty of detecting spurious correlations stems from the fact that they can not be identified based on observations alone (Section 2.2), and typically, some background knowledge is needed to judge the truthfulness of the correlations. In the tariff example, one needs to understand the nature of governance and the economic structures of the countries to be able to explain the underlying mechanisms that led to the correlation between tariff rate and economic growth. However, such background knowledge is not always straightforward. For a large exploratory

analysis, the analysts may discover pairs of relationships that interest them, but lack the necessary background knowledge to scrutinize these relationships. In this case, it is desirable to provide the analysts with some indication regarding the potential spuriousness (or the lack of) to help analyze relationships. This case study extends the visual analytics system in the previous case study to provide utilities to directly identify spurious correlations. The extended visual analytics system leverages Bayesian networks to generate causality hypothesis based on observations, and these hypothesis are matched against the correlations to identify potential spuriousness. Several visual analytics components are developed to help the analysts discover and inspect potential spurious correlations. The country-product correlation matrix view described in Section 5.2.1 is updated to highlight the spurious/non-spurious correlations indicated by the Bayesian network model. The causal path view is developed visualize the Bayesian network model to help analysts understand the rationale that leads to the indicated spuriousness/non-spuriousness. Several interactions are developed to allow adjustments of the Bayesian network model. The following sections will discuss the design and implementation of the extended visual analytics system, along with the details of each visual analytics component.

## 6.1 Design Requirements

One design challenge of this visual analytics system stems from the size and complexity of the data, as the analysts need to sift through a large set of detected correlations to discover the ones that interest them. At the same time, the analysts have to make sure the chosen correlations are not spurious. The sheer quantity of the detected correlations makes it unrealistic for the analysts to manually inspect each correlation for spuriousness; therefore, automatic algorithms are required. The Bayesian network model can generate causal hypotheses based on the data, which can be matched

against the detected correlations to identify spuriousness. These causal hypothesis are more reliable than correlations, but they still require additional inspection. This visual analytics system needs to facilitate the identification of spurious correlations through the usage of Bayesian networks, and, at the same time, provide utilities for inspection and manipulation of the Bayesian network models. Based on these discussions, several design requirements are summarized, and again, these are all related to the components of the proposed visual analytics framework in Chapter 3.

- R1** The system needs to provide an overview of detected correlations between the trade attributes and social stabilizes to assist the exploration on the correlations (Ch. 3: **C1, C3**).
- R2** The system needs to identify potential spurious correlations and highlight these spurious correlations (Ch. 3: **C1, C4**).
- R3** The system needs to reveal why some correlations are identified as spurious by the automatic model (Ch. 3: **C4, C5**).
- R4** The system needs to allow the analysts to adjust the output of the model through interactions (Ch. 3: **C6**).

## 6.2 Causality Modeling

One way to identify spurious correlations is to check the correlations against the edges in graphical causal models. These models can be conveniently trained using Bayesian networks, and many methods have been developed to extract causal relationships from data. As mentioned in Chapter 2, Bayesian networks are directed acyclic graphs (DAGs) that represent conditional dependencies (or the lack of) among variables, which can often be interpreted as causal relationships. To learn Bayesian

networks from data, there are two steps: structural learning and parameter estimation. The structure learning part in this case uses the *blip* library [175], which implements a score based approximation method [22]. The method searches the optimal structure by incrementally building the DAG using a greedy algorithm, where all possible structures are organized as a *k-tree* with bounded-treewidth, and the graph is built from the bottom up through searching for the parent sets that maximize the score. This allows the method to efficiently learn structures for very large networks, e.g. it takes about 20 seconds to train a network with about 3000 variables by setting the max treewidth to be 4, which is inconceivable for any standard structure learning algorithms. With the structure at hand, the parameters of the network can then be estimated with the standard MLE approach [69]. This Bayesian network learning method has two limitations that can influence our treatment on both the input and output: 1) The method only works with discrete Bayesian networks, so the input data needs to be discretized. The resulting Bayesian network will also only reflect the dependencies of the discretized data. 2) Bayesian networks learned from data are inherently ambiguous, and the structure learning algorithm will most likely settle to a local optimal; therefore, the result can only be treated as an approximation. These limitations impose some additional challenges on the interpretation and manipulation of the models, but the efficiency of the learning algorithm enables training on very large data, which is very suitable for this case study.

The Bayesian networks obtained from the the previously mentioned learning method reveal the dependency structure along with the conditional probability distribution (CPD) of each variable involved [21, 22]. However, the strength of each edge in the network is not directly available. This introduces a problem when one tries to interpret and visualize the edges as causal relations, as thinking about the strength of causal effects is an essential part of human reasoning. One way to compute the

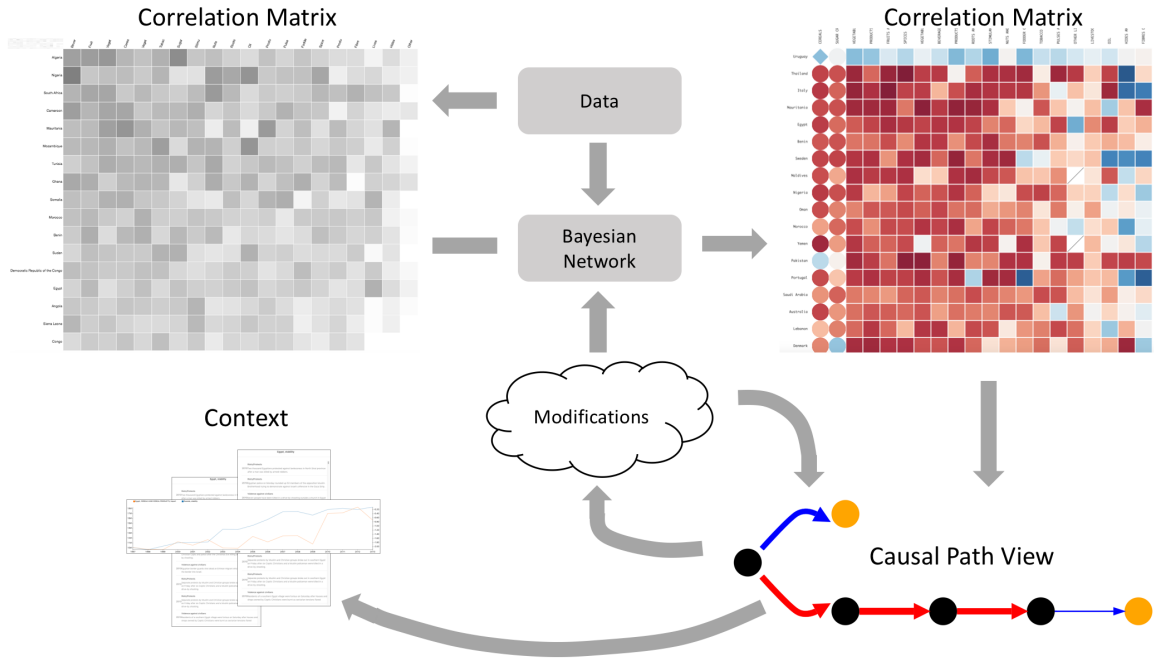


Figure 6.1: The System Workflow.

strength of causal relations is to use partial correlation [75], but it only works with continuous data; therefore, it can potentially introduce some discrepancy with the learning method used. This case study uses a method based on mutual information [176], where the strength of edge  $X \rightarrow Y$  is calculated as

$$w(X, Y) = \sum_{\mathbf{k} \in \Omega(\mathbf{Z})} P(\mathbf{k}) \sum_i P(X_i) \sum_j P(Y_j | X_i, \mathbf{k}) \log \frac{P(Y_j | X_i, \mathbf{k})}{P(Y_j | \mathbf{k})}$$

where  $\mathbf{Z} = \{Z \mid Z \in \text{parents of } Y \wedge Z \neq X\}$  and  $\Omega(\mathbf{Z})$  is all possible combinations of values of  $\mathbf{Z}$ . Note that this measure does not distinguish between positive and negative influence, thus the sign of  $w(X, Y)$  is estimated using the correlation between these two variables.

### 6.3 Visual Analytics Components

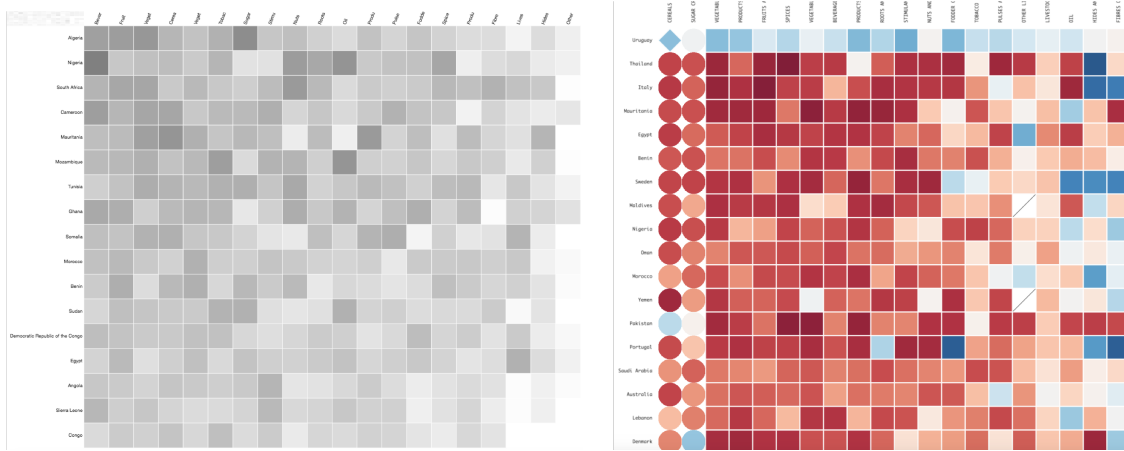
The goal of the visual analytics system is to extend the system in Chapter 5 to accommodate the requirements in Section 6.1. The extensions primarily serve to

support the identification and examination of spurious correlations. Figure 6.1 illustrates the workflow of the system. The extended visual analytics system leverages the Bayesian network model discussed in Section 6.2 to detect potential spurious correlations, and the country-product correlation matrix view in Section 5.2.1 is extended to include visual encoding that set apart the spurious correlations from the non-spurious ones. The benefit of having this view is to allow the analysts to quickly locate the spurious/non-spurious correlations of interest for further inspection (**R1**, **R2**). The analysts can further scrutinize the indicated spurious/non-spurious relationships by examining the causal path view, which displays the causal edges connected to the selected relations within the Bayesian network. This view reveals the rationale that leads to the indicated spuriousness/non-spuriousness, which helps the analysts to judge whether the indicated spuriousness/non-spuriousness is correct by inspecting the causal edges (**R3**). To support the analysts inspecting the causal edges, the system also provides some contextual information, such as the temporal trends of variables on both end of the causal relations and the ACLED events of the related countries (**R3**). The analysts are able to modify the causal models based on their judgments on the causal relations using a set of interaction provided by the system: removing/adding variables, removing causal edges, and setting the variable to a fixed value (**R4**). Upon reviewing the updated model, the analysts can make decisions on whether the selected correlation is truly spurious, and the correlation matrix can be updated accordingly. The following subsections discuss each visual analytics component in the system.

### 6.3.1 Correlation Matrix

The correlation matrix view in this case study is an extension to the country-product correlation matrix view described in Section 5.2.1. The previous country-





(a) The previous correlation matrix.




(b) The updated correlation matrix.

Figure 6.2: The Correlation Matrix.

product correlation matrix view (Figure 6.2(a)) visualizes the number of detected correlations between the trade attributes and the social stability measure for each country-product pair. However, many of the detected correlations may be spurious, which makes it less efficient to locate interesting country-product pairs. In this case study, the correlation matrix view (Figure 6.2(b)) incorporates the results from the Bayesian network model to highlight the potentially spurious correlations. This matrix view displays the correlation between a selected trade measure (import/export in this case study) and the social stability measure for each country-product pair. The rows in this matrix represent the countries, and the columns represent the food product categories. Each cell represents the correlation between the ordered pair  $(C, P, V)$  and the stability measure of the corresponding country, where  $C \in$  the set of all countries,  $P \in$  the set of all product categories, and  $V \in \{\text{import, export}\}$ . The Bayesian network model can be trained for every ordered pair  $(C, P, T)$ , together with the stability measure, but the analysts can also choose

a subset of the ordered pairs. The presence/absence of an edge between any ordered pair  $(C, P, T)$  and the stability measure in the Bayesian network model can be used to determine whether the correlation between the ordered pair and the stability measure is spurious, e.g. the absence of the edge indicates spuriousness. The information regarding spuriousness/non-spuriousness will be encoded in the matrix cells.

The cell in the matrix is colored in a diverging scale, with red representing negative correlations and blue representing positive correlations. The spurious/non-spurious of the correlations are encoded by the shape of cell:

- If the model indicates a true causal relationship between any of the ordered pair  $(C, P, T)$  and the stability measure, the corresponding cell in the matrix will be drawn as a  shape.
- If the model indicate an absence of causal relationship but correlation is detected, the correlation is labeled as spurious and the corresponding cell is drawn as a  shape.
- If the model can not determine if the correlation is spurious, the cell is drawn as a  shape.
- Cells with no detected correlations are drawn using the  $\diagup$  symbol.

The cells in the matrix are ordered in a way that the cells with non-spurious correlations are shifted to the top left corner, and the rest of the cells are ordered based on the magnitude of the correlations. The cells with larger correlations are shifted to the top left corner to be emphasized. The analysts can decide to order the cells based on either the positive or negative magnitude of the correlations. The ordering algorithm used here is the same as the one in Section 5.2.1 [167]. It is worth noting that the spuriousness/non-spuriousness indicated in the matrix can only be treated

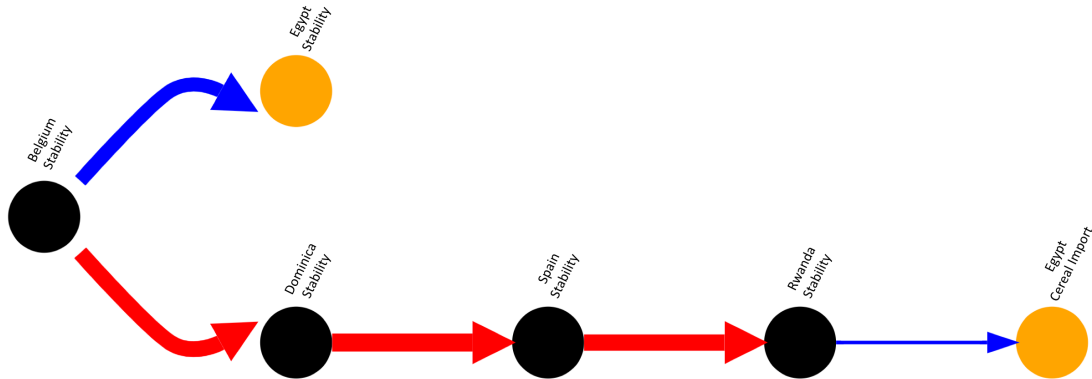


Figure 6.3: The Causal Path View.

as hints to guide further analysis, the analyst must investigate the rationale behind these indications, and this can be done by clicking on the cells of the matrix to bring up the causal path view.

### 6.3.2 Causal Path View

The causal path view (Figure 6.3) visualizes the Bayesian network (or the subset of) in a node-link diagram, in which the nodes represent the variables and the edges represent the causal relations. The Bayesian network trained from the data can potentially contain thousands of variables, which can create significant visual cluster if the entire network is displayed at once. The causal path view searches through the network and only displays the edges that are relevant to the selected correlation. This makes it easier to investigate the reason behind the indicated spuriousness/non-spuriousness of the correlation.

When the analysts click on a cell in the correlation matrix, it will bring up the the causal path view which draws the sub-network that contains the chains of the edges connected to the two variables of the corresponding correlation. These paths are extracted using a depth-first search starting from the two variables. When the correlation is indicated as non-spurious, the causal path view shows the path that goes

through the two variables of the correlation, along with the branches that connect to the path. This path reveals the causal edges related to the correlation, which provides the analysts some context to help understand the correlation. When the correlation is indicated as spurious, i.e. there is no causal edge connecting the two variables of the correlation, the causal path view may show the back-door path connecting the two variables. The back-door path is the path that introduces the confounding bias between the two variables, as discussed in Section 2.2. The existence of the back-door path may be the reason why the model thinks the detected correlation is not a consequence of causation. By examining the causal edges on this path, the analysts can better understand the rationale that leads to the indicated spuriousness, and subsequently make better judgments on the validity of the indications.

To help the analysts examine the causal edges, the causal path view uses several visual encodings to highlight the attributes of the causal edges. The width of the edge encodes the strength of the causal relation, where the strength is calculated using a method based on mutual information [176], as described in Section 6.2. The color of the edges indicate whether the causal influence is positive or negative, with red being negative and blue being positive. The sign of the causal inference is estimated using the sign of the correlation between the two variables. To make it easier to track the causal paths, the layout of the causal path view positions the nodes and edges in a way that emphasizes the flows of influences across the variables, by displaying the causal paths as flows starting from the left most nodes and ending at the right most nodes. This layout is computed using the Dagre Javascript library, which is based on a series of algorithms such as layered drawing [177] and edge-crossing minimization [170].

To illustrate the usage of the causal path view, an example is shown in Figure 6.3, which displays the back-door path between Egypt’s “cereal import quantity” and Egypt’s “social stability”. Referring back to the usage scenario shown in Sec-

tion 5.3.2, the analysts discovered a connection between Egypt’s wheat import and social stability. The analyst may be interested in inspecting whether the correlation between Egypt’s wheat import and social stability is spurious. In this case study, there is a detected correlation between Egypt’s “cereal import quantity” and Egypt’s “social stability”, as indicated by the correlation matrix. However, the Bayesian network model suggests a lack of causal relation between these two variables, which may possibly be explained by the presence of the back-door path. By closely examining back-door path, the analysts noticed Egypt’s “cereal import” and “social stability” are ultimately confounded by Belgium’s “stability”. The fact that Belgium’s “stability” positively influences Egypt’s “stability” may be justifiable given their close trade relationship. But the long path from Belgium’s “stability” to Egypt’s “cereal import” looks much more suspicious, especially given that the causal strength from Rwanda’s “stability” to Egypt’s “cereal import” is low. To inspect the validity of the causal relations, the analysts need more information, and some of them can be provided by the system.

### 6.3.3 *Context of Relationship*

The causal relationships shown in the causal path view can only be treated as an approximation, as discussed in Section 6.2. As such, it is necessary to check the validity of the edges in the causal path. This task is supported by providing some context of the causal relations. In this case study, the contextual information is visualized in two views: the trend comparison view and the event list.

#### **Trend Comparison**

By clicking on one edge in the causal path view, a trend comparison view will be displayed. The trend comparison view compares the time series of two related variables

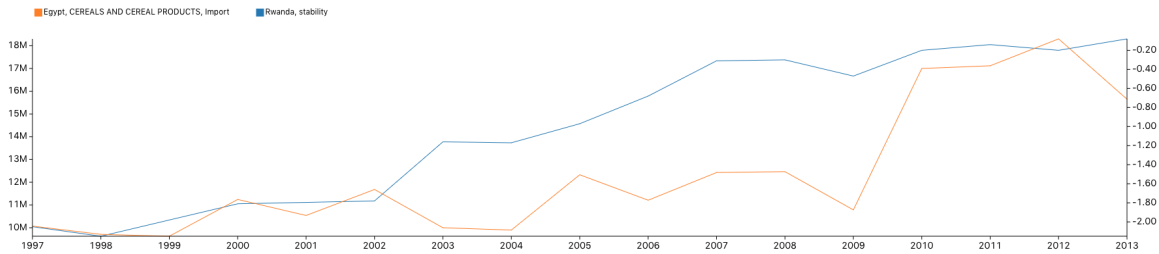


Figure 6.4: Trend Comparison.

to examine how these two variables are correlated. In this view, the time series of the two variables are displayed together with their scales adjusted to fit the same space. The analysts can examine how closely these two lines align with each other over different time period to determine how likely the correlation occurs by mere chance. For example, the trend comparison in Figure 6.4 shows that Rwanda’s stability has been improving consistently, which is not surprising considering Rwanda’s rapid economic growth in recent years. However, the trend of Egypt’s cereal import is less stable, although it does align well with Rwanda’s stability during the period between 2006 and 2011. Looking at the choropleth map view (described in Section 5.2.3), it can be shown that Egypt has no imports from Rwanda, while significant portion of Rwanda’s cereal import comes from Egypt. This may suggest that the causal relation from Rwanda’s stability to Egypt’s cereal import may not be true.

### Event List

Besides comparing the trends of two variables, it is also helpful to list the ACLED events of the countries related to the two variables. These lists of events can be displayed by brushing on the trend comparison view, and the lists contains events of the countries occurred within the time range selected by the brush, sorted by a chronological order. These events provide helpful information when the corresponding

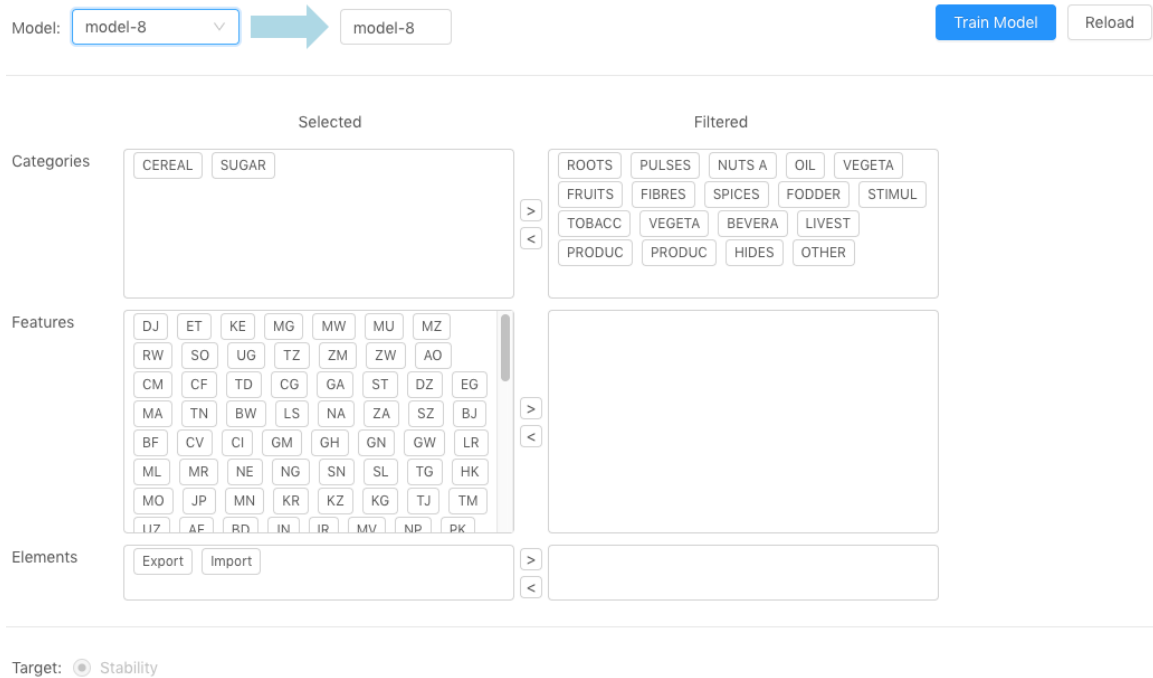


Figure 6.5: Variable Selection.

temporal trends are about social stability. The analysts may also find related events across the two lists to solidify their beliefs on the corresponding causal relations.

### 6.3.4 Model Tuning

Upon investigating the causal relations in the Bayesian network model, along with the context of relationships provided by the system, the analysts may find certain causal relations in the model disagreeable. And when such situation arises, it is natural for the analysts to demand the capability to modify the model in order to make causal relations in the model better aligned with the analysts' own judgments. Such modifications often alter the conditional probability distributions (CPDs) of the variables, which directly result in changes in causal strengths. Sometimes, these modifications may also alter the structure of the network, which subsequently changes

the indications on spuriousness/non-spuriousness in the correlation matrix view. This system provides three types of adjustments to the model:

- Variable Selection

It is very common for the analysts to have some preferences over what variables should be included in the Bayesian network model, either because they have some domain knowledge about the importance of variables, or because they want to limit the scope of the model to prevent the model from becoming overly complex to interpret. This variable selection process can take place before training the model, and it can also be performed after examining the existing model when the analysts feel the initial choices of variables are not optimal. Changing the choice of variables will result in a change in the optimal structure of the Bayesian network model. The system provides a tool to allow the analysts select the variables, as shown in Figure 6.5, in which the analysts are able to drag the tags to filter the countries, product categories, and import/export.

- Edge Deletion

When the analyst decides that an edge in the model is unreasonable, s/he can remove the edge, which subsequently changes the conditional probability distribution of the related variables. This modification may result in a change in causal strengths, as these causal strengths are calculated based on the CPDs (described in Section 6.2), and these changes in causal strengths are reflected in the causal path view.

- Value Selection

As for discrete Bayesian networks, it is possible to set a variable to a fixed value to observe its effects on the surrounding sub-network. Doing so effectively



removes the incoming causal edges to this variable, and it also changes the CPDs of the variables on which this variable has an influence, which subsequently changes the causal strengths of the related edges. The benefit of this operation is to enable the exploration on the effects of value assignments, which can help the analysts better understand the causal phenomenon in the model.

These model adjustments provide a means for the analysts to inject their domain knowledge into the model. By enabling interactions on the view and providing immediate feedback to these interactions, this visual analytics system allows the analysts to explore and experiment the effects of the model adjustments, which helps the analysts to develop a better understanding of the causal model. Based on these understandings, along with their own domain knowledge, the analysts can potentially make better judgments on the spuriousness/non-spuriousness of the correlations.

#### 6.4 Discussion

This case study develops a visual analytics system to assist the identification of spurious correlations during the exploration of a large spatiotemporal dataset. The Bayesian network model is incorporated for the automatic discovery of spurious correlations, and several visual analytics components are developed to assist the inspection of the model output. Interactions on the model is developed to assist the manipulation of the model as the means to inject domain knowledge into the model. These interactions also facilitate the understanding of the model output. Some contextual information is also provided to assist the inspection of the model. The understandings developed through the inspection of the model output can be used to justify/invalidate the indicated spuriousness of the correlations.

Identification of spurious correlation is a very difficult problem, and this case study does not seek a perfect solution to this problem. The proposed visual analytics

system only provides a means for hypothesis generation, and, in order to reliably identify spurious correlations, more rigorous scientific testing is required. It needs to be pointed out that this visual analytics system has not been evaluated based on usage scenarios or user studies. Future research can be directed to evaluating the effectiveness of this visual analytics systems at identifying spurious correlations. It is also notable that for causal hypothesis generation, changes in conditions over time cannot be ignored, and volatility in observational data can result in inconsistent causal models across different time steps. To address this issue, models that take into account temporality can be used, such as the dynamic Bayesian network model [178]. Visual analytics can also be used to tackle this issue; for example, Bayesian network models trained over different time steps can be visualized and compared to help develop understanding on the temporal evolution of models. These can all be potential future improvements for this visual analytics system.

## Chapter 7

### CONCLUSION

This thesis develops a visual analytics framework for correlation and causality analysis in which visual analytics methods are used to support a series of tasks including the exploration of data, feature engineering, correlation inspection, as well as examination and manipulation of causal models. Three case studies are conducted to demonstrate the effectiveness of this visual analytics framework. The first case study focuses on the exploration, linkage, and annotation of multiple media sources to explore drivers of discourse, in which a visual analytics based semantic matching scheme is used to extract relevant documents from media streams, and the Granger causality is applied to identify media drivers. The second case study focuses on enabling the users to understand how trade relationships impact local vulnerabilities over time in a global trade dataset, where visual analytics methods are integrated with correlation analysis and anomaly detection for pattern analysis and hypothesis generation. The third case study focuses on the identification of spurious correlations in which the visual analytics framework is used to facilitate exploration and manipulation of data and causal models. Each of these three case studies develops a visual analytics system for a domain specific problem guided by the proposed visual analytics framework, and the strengths and limitations of each system have been discussed. Currently, there are two notable limitations in these visual analytics systems. First, the semantic matching method proposed in the first case study is keyword based, which ignores the semantics capturable from the syntactic structures of sentences. Some natural language processing techniques such as the semantic role labeling [179] can be used to generate knowledge graphs from text documents [180], and these knowledge graphs

can be incorporated into a visual analytics system for semantic extraction in a large document collection. Second, the visual analytics system proposed in the third case study has not been evaluated for its effectiveness, and the evaluation can be done using a user study. Based on these discussions, my future work can be directed towards semantic extraction on text data and evaluations for the effectiveness of visual analytics on causality analysis.

## REFERENCES

- [1] T. Sternberg, “Chinese drought, bread and the Arab Spring,” *Applied Geography*, vol. 34, pp. 519–524, 2012.
- [2] G. Conti, J. Heckman, and S. Urzua, “The education-health gradient,” *American Economic Review*, vol. 100, no. 2, pp. 234–38, 2010.
- [3] L. B. Jennings, “Potential benefits of pet ownership in health promotion,” *Journal of Holistic Nursing*, vol. 15, no. 4, pp. 358–372, 1997.
- [4] R. A. Lewis and D. H. Reiley, “Online ads and offline sales: Measuring the effect of retail advertising via a controlled experiment on Yahoo!” *Quantitative Marketing and Economics*, vol. 12, no. 3, pp. 235–266, 2014.
- [5] J. Pearl, *Causality*. Cambridge University Press, 2009.
- [6] S. L. Bellamy, J. Y. Lin, and T. R. T. Have, “An introduction to causal modeling in clinical trials,” 2007.
- [7] J. Spahn, “Clinical trial efficacy: What does it really tell you?” *Journal of allergy and clinical immunology*, vol. 112, no. 5, pp. S102–S106, 2003.
- [8] R. DerSimonian and N. Laird, “Meta-analysis in clinical trials,” *Controlled clinical trials*, vol. 7, no. 3, pp. 177–188, 1986.
- [9] J. Pearl, “Why there is no statistical test for confounding, why many think there is, and why they are almost right,” *UCLA Department of Statistics Papers*, 1998.
- [10] J. Pearl *et al.*, “Causal inference in statistics: An overview,” *Statistics surveys*, vol. 3, pp. 96–146, 2009.
- [11] D. B. Rubin, “Estimating causal effects of treatments in randomized and non-randomized studies.” *Journal of Educational Psychology*, vol. 66, no. 5, p. 688, 1974.
- [12] D. T. Campbell and A. Erlebacher, “How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful,” *Compensatory education: A national debate*, vol. 3, pp. 185–210, 1970.
- [13] H. Wang, Y. Lu, S. T. Shutters, M. Steptoe, F. Wang, S. Landis, and R. Maciejewski, “A Visual Analytics Framework for Spatiotemporal Trade Network Analysis,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 331–341, 2018.
- [14] D. B. Stouffer, M. Sales-Pardo, M. I. Sirer, and J. Bascompte, “Evolutionary conservation of species’ roles in food webs,” *Science*, vol. 335, no. 6075, pp. 1489–1492, 2012.

- [15] J. Schaffer, “The metaphysics of causation,” in *The Stanford Encyclopedia of Philosophy*, fall 2016 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2016.
- [16] A. Falcon, “Aristotle on Causality,” in *The Stanford Encyclopedia of Philosophy*, spring 2019 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2019.
- [17] J. Neyman, “Sur les applications de la theorie des probabilités aux expériences agricoles: Essai des principes (In Polish). English translation by DM Dabrowska and TP Speed (1990),” *Statistical Science*, vol. 5, pp. 465–480, 1923.
- [18] J. Pearl, “Probabilistic reasoning in intelligent systems. 1988,” *San Mateo, CA: Kaufmann*, vol. 23, pp. 33–34.
- [19] I. Shpitser and J. Pearl, “Identification of conditional interventional distributions,” in *22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, 2006, pp. 437–444.
- [20] J. A. Gámez, J. L. Mateo, and J. M. Puerta, “Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood,” *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 106–148, 2011.
- [21] R. E. Neapolitan *et al.*, *Learning Bayesian Networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004, vol. 38.
- [22] M. Scanagatta, C. P. de Campos, G. Corani, and M. Zaffalon, “Learning bayesian networks with thousands of variables,” in *Advances in neural information processing systems*, 2015, pp. 1864–1872.
- [23] M. Joffe, M. Gambhir, M. Chadeau-Hyam, and P. Vineis, “Causal diagrams in systems epidemiology,” *Emerging themes in epidemiology*, vol. 9, no. 1, p. 1, 2012.
- [24] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins, “A structural approach to selection bias,” *Epidemiology*, vol. 15, no. 5, pp. 615–625, 2004.
- [25] N. Krieger and G. Davey Smith, “The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology,” *International journal of epidemiology*, vol. 45, no. 6, pp. 1787–1808, 2016.
- [26] D. Landuyt, S. Broekx, R. D’hondt, G. Engelen, J. Aertsens, and P. L. Goethals, “A review of bayesian belief networks in ecosystem service modelling,” *Environmental Modelling & Software*, vol. 46, pp. 1–11, 2013.
- [27] J. Y. Zhu, C. Zhang, H. Zhang, S. Zhi, V. O. Li, J. Han, and Y. Zheng, “pg-causality: Identifying spatiotemporal causal pathways for air pollutants with urban big data,” *IEEE Transactions on Big Data*, vol. 4, no. 4, pp. 571–585, 2017.

- [28] S. Andreassen, R. Hovorka, J. Benn, K. G. Olesen, and E. R. Carson, “A model-based approach to insulin adjustment,” in *Proceedings of the Third Conference on Artificial Intelligence in Medicine (AIME 91)*. Springer, 1991, pp. 239–248.
- [29] C. S. Jensen and A. Kong, “Blocking gibbs sampling for linkage analysis in large pedigrees with many loops,” *The American Journal of Human Genetics*, vol. 65, no. 3, pp. 885–901, 1999.
- [30] K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan, “Causal protein-signaling networks derived from multiparameter single-cell data,” *Science*, vol. 308, no. 5721, pp. 523–529, 2005.
- [31] H. Kim and Y. Park, “The impact of R&D collaboration on innovative performance in Korea: A Bayesian network approach,” *Scientometrics*, vol. 75, no. 3, p. 535, 2008.
- [32] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.
- [33] M. Eichler, “Causal inference in time series analysis,” *Causality: statistical perspectives and applications*. Wiley, Chichester, pp. 327–354, 2012.
- [34] A. Arnold, Y. Liu, and N. Abe, “Temporal causal modeling with graphical granger methods,” in *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2007, pp. 66–75.
- [35] A. Inselberg and B. Dimsdale, “Parallel coordinates: a tool for visualizing multi-dimensional geometry,” in *Proceedings of the First IEEE Conference on Visualization: Visualization ‘90*, Oct 1990, pp. 361–378.
- [36] Z. Zhang, K. T. McDonnell, and K. Mueller, “A network-based interface for the exploration of high-dimensional data spaces,” in *2012 IEEE Pacific Visualization Symposium*, Feb 2012, pp. 17–24.
- [37] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller, “Visual correlation analysis of numerical and categorical data on the correlation map,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 2, pp. 289–303, Feb 2015.
- [38] H. Qu, W.-Y. Chan, A. Xu, K.-L. Chung, K.-H. Lau, and P. Guo, “Visual Analysis of the Air Pollution Problem in Hong Kong,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1408–1415, 2007.
- [39] J. Xia, W. Chen, Y. Hou, W. Hu, X. Huang, and D. S. Ebertk, “Dimscanner: A relation-based visual exploration approach towards data dimension inspection,” in *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2016, pp. 81–90.

- [40] X. Yuan, D. Ren, Z. Wang, and C. Guo, “Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2625–2633, 2013.
- [41] T. N. Dang, A. Anand, and L. Wilkinson, “Timeseer: Scagnostics for high-dimensional time series,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 3, pp. 470–483, 2012.
- [42] T.-Y. Lee and H.-W. Shen, “Visualization and exploration of temporal trend relationships in multivariate time-varying data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1359–1366, 2009.
- [43] B. Schneider, M. Carnoy, J. Kilpatrick, W. H. Schmidt, and R. J. Shavelson, *Estimating causal effects using experimental and observational design*. American Educational & Research Association, 2007.
- [44] H. Reichenbach, *The direction of time*. Univ of California Press, 1991, vol. 65.
- [45] P. Suppes, “A Probabilistic Theory of Causality,” *British Journal for the Philosophy of Science*, vol. 24, no. 4, pp. 409–410, 1973.
- [46] N. Cartwright, “Causal laws and effective strategies,” *Noûs*, pp. 419–437, 1979.
- [47] C. Hitchcock, “Probabilistic Causation,” in *The Stanford Encyclopedia of Philosophy*, fall 2018 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018.
- [48] S. Greenland, J. M. Robins, J. Pearl *et al.*, “Confounding and collapsibility in causal inference,” *Statistical science*, vol. 14, no. 1, pp. 29–46, 1999.
- [49] J. Berkson, “Limitations of the application of fourfold table analysis to hospital data,” *Biometrics Bulletin*, vol. 2, no. 3, pp. 47–53, 1946.
- [50] J. Pearl, “[Bayesian analysis in expert systems]: comment: graphical models, causality and intervention,” *Statistical Science*, vol. 8, no. 3, pp. 266–269, 1993.
- [51] D. Borsboom, G. J. Mellenbergh, and J. Van Heerden, “The theoretical status of latent variables.” *Psychological review*, vol. 110, no. 2, p. 203, 2003.
- [52] R. M. Baron and D. A. Kenny, “The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.” *Journal of personality and social psychology*, vol. 51, no. 6, p. 1173, 1986.
- [53] C. M. Judd and D. A. Kenny, “Process analysis: Estimating mediation in treatment evaluations,” *Evaluation review*, vol. 5, no. 5, pp. 602–619, 1981.
- [54] L. R. James and J. M. Brett, “Mediators, moderators, and tests for mediation.” *Journal of Applied Psychology*, vol. 69, no. 2, p. 307, 1984.



- [55] J. Pearl, “Interpretation and identification of causal mediation.” *Psychological methods*, vol. 19, no. 4, p. 459, 2014.
- [56] —, “Causal diagrams for empirical research,” *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.
- [57] P. Spirtes, C. Glymour, R. Scheines, C. Meek, S. Fienberg, and E. Slate, “Prediction and experimental design with graphical causal models,” *Computation, causation, and discovery*, pp. 65–94, 1999.
- [58] B. C. Sauer, M. A. Brookhart, J. Roy, and T. VanderWeele, “A review of covariate selection for non-experimental comparative effectiveness research,” *Pharmacoepidemiology and drug safety*, vol. 22, no. 11, pp. 1139–1145, 2013.
- [59] Y. Liu, Q. Du, Q. Wang, H. Yu, J. Liu, Y. Tian, C. Chang, and J. Lei, “Causal inference between bioavailability of heavy metals and environmental factors in a large-scale region,” *Environmental pollution*, vol. 226, pp. 370–378, 2017.
- [60] L. A. Sierra, V. Yepes, T. García-Segura, and E. Pellicer, “Bayesian network method for decision-making about the social sustainability of infrastructure projects,” *Journal of Cleaner Production*, vol. 176, pp. 521–534, 2018.
- [61] M. B. Sesen, A. E. Nicholson, R. Banares-Alcantara, T. Kadir, and M. Brady, “Bayesian networks for clinical decision support in lung cancer care,” *Plos One*, vol. 8, no. 12, p. e82349, 2013.
- [62] A. Bhattacharjee, “Application of bayesian approach in cancer clinical trial,” *World journal of oncology*, vol. 5, no. 3, p. 109, 2014.
- [63] J. Pearl and S. Russell, “Bayesian networks,” *UCLA Department of Statistic Papers*.
- [64] D. M. Chickering, D. Heckerman, and C. Meek, “Large-sample learning of Bayesian networks is NP-hard,” *Journal of Machine Learning Research*, vol. 5, no. Oct, pp. 1287–1330, 2004.
- [65] T. Verma and J. Pearl, “Equivalence and synthesis of causal models,” in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. Elsevier Science Inc., 1990, pp. 255–270.
- [66] —, “An algorithm for deciding if a set of observed independencies has a causal explanation,” in *Uncertainty in artificial intelligence*. Elsevier, 1992, pp. 323–330.
- [67] D. Dash and M. J. Druzdzel, “Robust independence testing for constraint-based learning of causal structure,” in *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 167–174.

- [68] R. Opgen-Rhein and K. Strimmer, “From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data,” *BMC systems biology*, vol. 1, no. 1, p. 37, 2007.
- [69] D. Heckerman, “A tutorial on learning with Bayesian networks,” in *Learning in graphical models*. Springer, 1998, pp. 301–354.
- [70] N. Elmqvist and P. Tsigas, “Growing squares: Animated visualization of causal relations,” in *Proceedings of the 2003 ACM Symposium on Software Visualization*. ACM, 2003, pp. 17–ff.
- [71] K. A. Baker, P. C. Fishburn, and F. S. Roberts, “Partial orders of dimension 2,” *Networks*, vol. 2, no. 1, pp. 11–28, 1972.
- [72] N. Elmqvist and P. Tsigas, “Causality visualization using animated growing polygons,” in *IEEE Symposium on Information Visualization 2003*. IEEE, 2003, pp. 189–196.
- [73] N. Kadaba, P. Irani, and J. Leboe, “Visualizing causal semantics using animations,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1254–1261, 2007.
- [74] J.-D. Zapata-Rivera, E. Neufeld, and J. E. Greer, “Visualization of Bayesian belief networks,” in *Proceedings of IEEE Visualization*, 1999, pp. 85–88.
- [75] J. Wang and K. Mueller, “The visual causality analyst: An interactive interface for causal reasoning,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 230–239, Jan 2016.
- [76] T. M. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.
- [77] J. Wang and K. Mueller, “Visual causality analysis made practical,” in *IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2017, pp. 151–161.
- [78] J. Y. Zhu, C. Sun, and V. O. K. Li, “Granger-causality-based air quality estimation with spatio-temporal (s-t) heterogeneous big data,” in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2015, pp. 612–617.
- [79] J. C. Roberts, “State of the art: Coordinated & multiple views in exploratory visualization,” in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*. IEEE, 2007, pp. 61–71.
- [80] Y. Liu, S. Barlowe, Y. Feng, J. Yang, and M. Jiang, “Evaluating exploratory visualization systems: A user study on how clustering-based visualization systems support information seeking from large document collections,” *Information Visualization*, vol. 12, no. 1, pp. 25–43, 2013.

- [81] B. Shipley, *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press, 2016.
- [82] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park, "ivisclustering: An interactive visual document clustering via topic modeling," in *Computer Graphics Forum*, vol. 31, no. 3pt3. Wiley Online Library, 2012, pp. 1155–1164.
- [83] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 543–552.
- [84] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "Deadline: Interactive visual analysis of text data through event identification and exploration," in *IEEE Conference on Visual Analytics Science and Technology (VAST), 2012*. IEEE, 2012, pp. 93–102.
- [85] W. Cui, S. Liu, Z. Wu, and H. Wei, "How hierarchical topics evolve in large text corpora," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2281–2290, 2014.
- [86] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, Jan 2002.
- [87] F. Wanner, W. Jentner, T. Schreck, A. Stoffel, L. Sharalieva, and D. A. Keim, "Integrated visual analysis of patterns in time series and text data-workflow and application to financial data analysis," *Information Visualization*, vol. 15, no. 1, pp. 75–90, 2016.
- [88] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: Automatic generation of annotated stock visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2707–2716.
- [89] S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker, "Exploring topical lead-lag across corpora," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 115–129, Jan 2015.
- [90] Y. Lu, M. Steptoe, S. Burke, H. Wang, J.-Y. Tsai, H. Davulcu, D. Montgomery, S. R. Corman, and R. Maciejewski, "Exploring evolving media discourse through event cueing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 220–229, 2016.
- [91] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.
- [92] D. Metzler, S. Dumais, and C. Meek, *Similarity measures for short segments of text*. Springer, 2007.

- [93] R. Mihalcea, C. Corley, and C. Strapparava, “Corpus-based and knowledge-based measures of text semantic similarity,” in *Association for the Advancement of Artificial Intelligence*, vol. 6, 2006, pp. 775–780.
- [94] R. Navigli, “Word sense disambiguation: A survey,” *ACM Computing Surveys (CSUR)*, vol. 41, no. 2, p. 10, 2009.
- [95] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.
- [96] G. A. Miller, “WordNet: A lexical database for English,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [97] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit.” in *ACL (System Demonstrations)*, 2014, pp. 55–60.
- [98] S. C. Leman, L. House, D. Maiti, A. Endert, and C. North, “Visual to parametric interaction (v2pi),” *Plos One*, vol. 8, no. 3, p. e50474, 2013.
- [99] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, “Visualizing the non-visual: spatial analysis and interaction with information from text documents,” in *Proceedings of Information Visualization*. IEEE, 1995, pp. 51–58.
- [100] T. Sørensen, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons,” *Biologiske Skrifter*, vol. 5, pp. 1–34, 1948.
- [101] M. Bostock, V. Ogievetsky, and J. Heer, “D<sup>3</sup> data-driven documents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.
- [102] A. Endert, L. Bradel, and C. North, “Beyond control panels: Direct manipulation for visual analytics,” *IEEE Computer Graphics & Applications*, vol. 33, no. 4, pp. 6–13, July 2013.
- [103] A. Endert, P. Fiaux, and C. North, “Semantic interaction for visual text analytics,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2012, pp. 473–482.
- [104] A. Endert, S. Fox, D. Maiti, and C. North, “The semantics of clustering: analysis of user-generated spatializations of text documents,” in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 555–562.
- [105] H. B. Asher, *Causal Modeling*. Sage, 1983, no. 3.
- [106] F. Diebold, *Elements of Forecasting*. South-Western College Pub., 1998.

- [107] J. Fulda, M. Brehmer, and T. Munzner, “Timelinecurator: Interactive authoring of visual timelines from unstructured text,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 300–309, 2016.
- [108] T. Gao, J. R. Hullman, E. Adar, B. Hecht, and N. Diakopoulos, “Newsviews: An automated pipeline for creating custom geovisualizations for news,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2014, pp. 3005–3014.
- [109] J. Zhao, M. Glueck, S. Breslav, F. Chevalier, and A. Khan, “Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations,” *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2016.
- [110] C. Raleigh, A. Linke, H. Hegre, and J. Karlsen, “Introducing ACLED: An armed conflict location and event dataset,” *Journal of Peace Research*, vol. 47, no. 5, pp. 651–660, 2010.
- [111] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, “Pair analytics: Capturing reasoning processes in collaborative visual analytics,” in *44th Hawaii International Conference on System Sciences*. IEEE, 2011, pp. 1–10.
- [112] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
- [113] S. C. Herring, A. Hoell, M. P. Hoerling, J. P. Kossin, C. J. Schreck, and P. A. Stott, “Explaining extreme events of 2015 from a climate perspective,” *Bulletin of the American Meteorological Society*, vol. 97, p. Sii, 2016.
- [114] H. Buhaug, “Climate change and conflict: Taking stock,” *Peace Economics, Peace Science and Public Policy*, vol. 22, pp. 331–338, 2016.
- [115] O. M. Theisen, H. Holtermann, and H. Buhaug, “Climate wars? assessing the claim that drought breeds conflict,” *MIT Press*, 2011.
- [116] G. K. MacDonald, K. A. Brauman, S. Sun, K. M. Carlson, E. S. Cassidy, J. S. Gerber, and P. C. West, “Rethinking agricultural trade relationships in an era of globalization,” *BioScience*, p. biu225, 2015.
- [117] J. J. McCarthy, *Climate change 2001: impacts, adaptation, and vulnerability: contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2001.
- [118] M. G. Marshall, “Fragility, instability, and the failure of states,” Council on Foreign Relations, Washginton, DC, Tech. Rep., October 2008. [Online]. Available: <https://www.cfr.org/report/fragility-instability-and-failure-states>
- [119] S. M. Hsiang, K. C. Meng, and M. A. Cane, “Civil conflicts are associated with the global climate,” *Nature*, vol. 476, no. 7361, pp. 438–441, 2011.

- [120] United States Department of Defense, “Quadrennial Defense Review Report,” Washington, DC, 2014. [Online]. Available: [http://archive.defense.gov/pubs/2014\\_Quadrennial\\_Defense\\_Review.pdf](http://archive.defense.gov/pubs/2014_Quadrennial_Defense_Review.pdf)
- [121] D. Garlaschelli and M. I. Loffredo, “Structure and evolution of the world trade network,” *Physica A: Statistical Mechanics and its Applications*, vol. 355, no. 1, pp. 138–144, Sep. 2005.
- [122] L. De Benedictis and L. Tajoli, “Comparing Sectoral International Trade Networks,” *Aussenwirtschaft*, no. 65, pp. 167–189, 2010.
- [123] P. Kaluza, A. Klzsch, M. T. Gastner, and B. Blasius, “The complex network of global cargo ship movements,” *Journal of The Royal Society Interface*, p. rsif20090495, Jan. 2010.
- [124] L. De Benedictis and L. Tajoli, “The world trade network,” *The World Economy*, vol. 34, no. 8, pp. 1417–1454, 2011.
- [125] M. Konar, C. Dalin, S. Suweis, N. Hanasaki, A. Rinaldo, and I. Rodriguez-Iturbe, “Water for food: The global virtual water trade network,” *Water Resources Research*, vol. 47, no. 5, pp. W05 520:1–17, May 2011.
- [126] C. Dalin, M. Konar, N. Hanasaki, A. Rinaldo, and I. Rodriguez-Iturbe, “Evolution of the global virtual water trade network,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5989–5994, Apr. 2012.
- [127] S. Suweis, M. Konar, C. Dalin, N. Hanasaki, A. Rinaldo, and I. Rodriguez-Iturbe, “Structure and controls of the global virtual water trade network,” *Geophysical Research Letters*, vol. 38, no. 10, p. L10403, May 2011.
- [128] D. Garlaschelli and M. I. Loffredo, “Fitness-Dependent Topological Properties of the World Trade Web,” *Physical Review Letters*, vol. 93, no. 18, p. 188701, Oct. 2004.
- [129] D. Garlaschelli, T. D. Matteo, T. Aste, G. Caldarelli, and M. I. Loffredo, “Interplay between topology and dynamics in the World Trade Web,” *The European Physical Journal B*, vol. 57, no. 2, pp. 159–164, May 2007.
- [130] M. Barigozzi, G. Fagiolo, and G. Mangioni, “Identifying the community structure of the international-trade multi-network,” *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 11, pp. 2051–2066, Jun. 2011.
- [131] P. Shi, J. Zhang, B. Yang, and J. Luo, “Hierarchicality of Trade Flow Networks Reveals Complexity of Products,” *Plos One*, vol. 9, no. 6, p. e98247, Jun. 2014.
- [132] W. Luo, P. Yin, Q. Di, F. Hardisty, and A. M. MacEachren, “A geovisual analytic approach to understanding geo-social relationships in the international trade network,” *Plos One*, vol. 9, no. 2, p. e88666, 2014.

- [133] Y. Lu, H. Wang, S. Landis, and R. Maciejewski, “A visual analytics framework for identifying topic drivers in media events,” *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2018.
- [134] G. Fagiolo and M. Mastrorillo, “Does human migration affect international trade? A complex-network perspective,” *Plos One*, vol. 9, no. 5, p. e97331, 2014.
- [135] K. Kanemoto, D. Moran, M. Lenzen, and A. Geschke, “International trade undermines national emission reduction targets: New evidence from air pollution,” *Global Environmental Change*, vol. 24, pp. 52–59, 2014.
- [136] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne, “Flowstrates: An approach for visual exploration of temporal origin-destination data,” in *Computer Graphics Forum*, vol. 30, no. 3. Wiley Online Library, 2011, pp. 971–980.
- [137] D. Guo and X. Zhu, “Origin-Destination Flow Data Smoothing and Mapping,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2043–2052, 2014.
- [138] R. Scheepens, N. Willems, H. van de Wetering, and J. van Wijk, “Interactive Density Maps for Moving Objects,” *IEEE Computer Graphics and Applications*, vol. 32, no. 1, pp. 56–66, 2012.
- [139] R. Scheepens, H. van de Wetering, and J. J. van Wijk, “Contour based visualization of vessel movement predictions,” *International Journal of Geographical Information Science*, vol. 28, no. 5, pp. 891–909, 2014.
- [140] J. Chae, D. Thom, H. Bosch, J. Yang, R. Maciejewski, D. S. Ebert, and T. Ertl, “Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition,” in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2012.
- [141] J. Chae, Y. Cui, Y. Jang, G. Wang, A. Malik, and D. S. Ebert, “Trajectory-based Visual Analytics for Anomalous Human Movement Analysis using Social Media,” in *EuroVis Workshop on Visual Analytics (EuroVA)*, E. Bertini and J. C. Roberts, Eds. The Eurographics Association, 2015.
- [142] S. T. Shutters and R. Muneeppeerakul, “Agricultural Trade Networks and Patterns of Economic Development,” *Plos One*, vol. 7, no. 7, pp. e39 756:1–9, Jul. 2012.
- [143] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, “Superfamilies of evolved and designed networks,” *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.
- [144] L. Zhang, G. Qian, and L. Zhang, “Network motif & triad significance profile analyses on software system,” *W. Trans. on Comp.*, vol. 7, no. 6, pp. 756–765, 2008.

- [145] “Food and Agriculture Organization of the United Nations Statistics Division,” <http://www.fao.org/faostat/en/#home>, 2015.
- [146] D. Kaufmann, A. Kraay, and M. Mastruzzi, “The worldwide governance indicators: methodology and analytical issues,” *Hague Journal on the Rule of Law*, vol. 3, no. 2, pp. 220–246, 2011.
- [147] J. A. Reyes and R. Kali, “The Architecture of Globalization: A Network Approach to International Economic Integration,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 922059, May 2006.
- [148] G. Fagiolo, J. Reyes, and S. Schiavo, “The evolution of the world trade web: a weighted-network analysis,” *Journal of Evolutionary Economics*, vol. 20, no. 4, pp. 479–514, Aug. 2010.
- [149] J. Reyes, S. Schiavo, and G. Fagiolo, “Using complex network analysis to assess the evolution of international economic integration: The cases of East Asia and Latin America,” LEM Working Paper Series, Working Paper 2007/25, 2007.
- [150] D. A. Keim, F. Mansmann, and J. Thomas, “Visual Analytics: How Much Visualization and How Much Analytics?” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 2, pp. 5–8, 2010.
- [151] R. A. Hanneman and M. Riddle, *Introduction to social network methods*. Riverside, California, USA: University of Riverside, 2005.
- [152] K. Faust, “Very local structure in social networks,” *Sociological Methodology*, vol. 37, pp. 209–256, 2007.
- [153] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, ser. Structural Analysis in the Social Sciences. New York: Cambridge University Press, 1994.
- [154] D. Cartwright and F. Harary, “Structural balance: a generalization of heider’s theory,” *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956.
- [155] P. W. Holland and S. Leinhardt, “Local structure in social networks,” *Sociological Methodology*, vol. 7, pp. 1–45, 1976.
- [156] J. Yoon, S. R. Thye, and E. J. Lawler, “Exchange and cohesion in dyads and triads: A test of simmels hypothesis,” *Social science research*, vol. 42, no. 6, pp. 1457–1466, 2013.
- [157] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network Motifs: Simple Building Blocks of Complex Networks,” *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.
- [158] C. Prell and J. Skvoretz, “Looking at social capital through triad structures,” *Connections*, vol. 28, no. 2, pp. 4–16, 2008.



- [159] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius, “The complex network of global cargo ship movements,” *Journal of The Royal Society Interface*, 2010.
- [160] J. S. Waters and J. H. Fewell, “Information processing in social insect networks,” *Plos One*, vol. 7, no. 7, p. e40337, 2012.
- [161] M. Szell and S. Thurner, “Measuring social dynamics in a massive multiplayer online game,” *Social Networks*, vol. 32, no. 4, pp. 313–329, 2010.
- [162] G. Facchetti, G. Iacono, and C. Altafini, “Computing global structural balance in large-scale signed social networks,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 52, pp. 20 953–20 958, 2011.
- [163] L. Bargigli, G. di Iasio, L. Infante, F. Lillo, and F. Pierobon, “The multiplex structure of interbank networks,” *Quantitative Finance*, vol. 15, no. 4, pp. 673–691, 2015.
- [164] Z. Moaz, *Networks of nations: The evolution, structure, and impact of international networks, 1816-2001*, ser. Structural analysis in the social sciences. New York: Cambridge University Press, 2011, vol. 32.
- [165] S. C. Lee, R. G. Muncaster, and D. A. Zinnes, “The friend of my enemy is my enemy: Modeling triadic international relationships,” *Synthese*, vol. 100, no. 3, pp. 333–358, 1994.
- [166] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, “Intensity and coherence of motifs in weighted complex networks,” *Physical Review E*, vol. 71, no. 6, pp. 065 103:1–4, 2005.
- [167] E. Mäkinen and H. Siirtola, “Reordering the reorderable matrix as an algorithmic problem,” in *International Conference on Theory and Application of Diagrams*. Springer, 2000, pp. 453–468.
- [168] S. Makridakis and M. Hibon, “ARMA models and the Box–Jenkins methodology,” *Journal of Forecasting*, vol. 16, no. 3, pp. 147–163, 1997.
- [169] J. S. Hunter, “The exponentially weighted moving average,” *Journal of quality technology*, vol. 18, no. 4, pp. 203–210, 1986.
- [170] M. Jünger and P. Mutzel, “2-layer straightline crossing minimization: Performance of exact and heuristic algorithms,” in *Graph Algorithms And Applications I*. World Scientific, 2002, pp. 3–27.
- [171] F. Franzetti, A. Pezzoli, and M. Baliani, *Rethinking Water Resources Management Under a Climate Change Perspective: From National to Local Level. The Case of Thailand*. Cham, Switzerland: Springer, 2017, pp. 169–195.
- [172] R. Satloff and D. Schenker, “Contingency Planning Memorandum No. 19: Political Instability in Jordan,” Council on Foreign Relations, Washington, DC, Tech. Rep., May 2013. [Online]. Available: <http://www.cfr.org/jordan/political-instability-jordan/p30698>

- [173] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, “A taxonomy and survey of dynamic graph visualization,” in *Computer Graphics Forum*, vol. 36, no. 1. Wiley Online Library, 2017, pp. 133–159.
- [174] D. A. Irwin, “Interpreting the tariff-growth correlation of the late 19th century,” *American Economic Review*, vol. 92, no. 2, pp. 165–169, 2002.
- [175] M. Scanagatta, “blip.” [Online]. Available: <https://github.com/mauro-idsia/blip>
- [176] A. E. Nicholson and N. Jitnah, “Using mutual information to determine relevance in Bayesian networks,” in *Pacific Rim international conference on artificial intelligence*. Springer, 1998, pp. 399–410.
- [177] E. R. Gansner, E. Koutsofios, S. C. North, and K.-P. Vo, “A technique for drawing directed graphs,” *IEEE Transactions on Software Engineering*, vol. 19, no. 3, pp. 214–230, 1993.
- [178] P. Dagum, A. Galper, and E. Horvitz, “Dynamic network models for forecasting,” in *Proceedings of the eighth international conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1992, pp. 41–48.
- [179] M. Palmer, D. Gildea, and N. Xue, “Semantic role labeling,” *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–103, 2010.
- [180] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, and T. Bogaard, “Building event-centric knowledge graphs from news,” *Journal of Web Semantics*, vol. 37, pp. 132–151, 2016.